Preparing Scottish census data in I-CeM for the *British Business Census of Entrepreneurs* (BBCE)

Harry Smith, Carry van Lieshout, Piero Montebruno, and Robert J. Bennett

hjs57@cam.ac.uk cv313@cam.ac.uk pfm27@cam.ac.uk rjb7@cam.ac.uk

Working Paper 20:
Working paper series from ESRC project ES/M010953:
Drivers of Entrepreneurship and Small Businesses

University of Cambridge, Department of Geography and Cambridge Group for the History of Population and Social Structure, Downing Place, Cambridge, CB2 3EN, UK.

September 2019

Comments are welcomed on this paper: contact the authors as above.

Keywords: Entrepreneurship, Employers, Self-employment, Small businesses, Census

JEL Codes: L26, L25, D13, D22

Preparing Scottish census data in I-CeM for the *British Business Census of Entrepreneurs* (BBCE)

Harry Smith, Carry van Lieshout, Piero Montebruno, and Robert J. Bennett

Working Paper 20: ESRC project ES/M010953: Drivers of Entrepreneurship and Small Businesses, University of Cambridge

1. Introduction

This paper discusses how the Scottish census data on employer and own account business proprietors contained in the population censuses for 1851-1901 were prepared for analysis and subsequent deposit as part of the *British Business Census of Entrepreneurs* (BBCE) data deposit at UK Data Archive (UKDA). For BBCE, the transcripts of the censuses, and coding of individuals, are mostly derived from the UKDA data deposit of *The Integrated Census Microdata* (I-CeM). BBCE adds to I-CeM identification and coding of entrepreneurs, plus data enrichment, corrections and infills of those missing in I-CeM. The BBCE and I-CeM can be linked through the individual identifiers for each entrepreneur identified in the censuses to provide a complete resource of the census information on entrepreneurs. The Scottish data were prepared using a special financial supplementation of the ESRC-supported project ES/M010953 'Drivers of Entrepreneurship and Small Businesses'. This recognised the importance of expanding the data available on England and Wales to give full aggregates for Britain as a whole.

The methodology used to prepare the data for Scotland largely follow that developed for the ESRC project covering England and Wales, but had to be adapted in various respects. This paper outlines the methods used and the decisions made in the Scottish database construction. It also acts as introduction to the Scottish aspects of the documentation of the BBCE database deposited at UKDA, which are defined in the *BBCE User Guide*. The methods used for England and Wales are described in other Working Papers listed at the end of this Working

Paper, to which the reader is referred for more detail on the development of the methodologies employed.

The paper describes how the information for Scotland has been brought up to the level available in BBCE for England and Wales on an aligned and fully compatible basis. The full coding resulting from the Scottish efforts discussed here is included in the BBCE. Additionally, downloads are provided with this Working Paper 20 of the reconstruction decision process, cut-off information for farms by parish, and the estimation of the workers in the whole active population with their occodes. This allows any user to reconstruct the same estimates as we have developed based on the same data (with warnings to users on the preliminary nature of some of the data noted). The Scottish Burghs classification is given in a download with WP 6. Downloads of the reconstruction model estimates and a guide to how the estimation process works are provided with WP 19.

1.1 Scottish Census administration

The data used derives from the population censuses. These were administered for England and Wales by the General Register Office (GRO). In Scotland the census derived from the same general legislation and was administered by the GRO for the whole of Britain in 1851. However, from 1861 although under the same GRO framework in London, the administration passed to Scotland under its own Registrar General. Thus, from 1861, the Scottish censuses differed in some respects from the England and Wales censuses: there were differences in administration, minor changes of instructions to householders, some additional questions, and some differences in arrangement of published tables and the geographical units for which results were published that have some effects on data construction and the comparisons that can now be made.

An important aspect was that for 1851 the census was administered and tabulated for *Civil* Counties and Divisions or Parishes, but from 1861 onwards *Registration* Districts and Counties were used as well as the *Civil* units. The differences between the areas covered by Civil and Registration units make comparisons with some of the data available now more difficult. Some of these differences are large at parish level, and some affect counties as well.

Although 'a copious index' of the relation between the two was published, not all comparisons that can now be made between the different units are precise.[1]

It should be noted that *Counties in I-CeM are coded on Civil County* boundaries, but the publications of the census for Scotland varied. All our calculations for the Scottish data and in BBCE at county level relate to these I-CeM codes. Hence *the Scottish BBCE is based on Civil County boundaries*, where relevant. Any comparisons that users of the data make will need to note that the published census tables will not match precisely if GRO used Registration Counties. However, because of major changes to County boundaries in the 1890s the civil and registration counties were much closer for 1901 and 1911.

As in England and Wales the data now available derives from the original Census Enumerators Books (CEBs). In Scotland, from 1861 the enumerator passed the CEBs and householders' returns to local Sheriff-Clerks, who then passed to Registrars, and finally to the Scottish Registrar General's Head Office.

1.2 Scotland's I-CeM data

The CEBs from the GRO administrative system are now the data that have become available as a digital database through the ESRC project based at the University of Essex which deposited the database at the UKDA: *The Integrated Census Microdata* (I-CeM).[2] The original records are held at The National Archives (TNA) for England and Wales, but in Scotland are held by The National Records of Scotland and accessed through *Scotland's People*. The records in I-CeM were originally transcribed by FindMyPast (FMP) for 1851-71 and 1891-1901; and by the Genealogical Society of Utah (GSU) for 1881. The I-CeM team turned these genealogical sources into a generic database structure, adding coding for occupations, household structure, locations, and other features: for which, see the *I-CeM User Guide*.[3] An important difference from England and Wales I-CeM is that 1911 is not available for Scotland, although the data are available at *Scotland's People.*

[1] *Census of Scotland, 1861. Population tables and report, Vol. II*, *Parliamentary Papers*, 51 (1864), lxii.
[2] Higgs, Edward and Schürer, Kevin (University of Essex) (2014) *The Integrated Census Microdata (I-CeM)* UKDA, SN-7481
[3] https://www.essex.ac.uk/history/research/icem/documents/icem_guide.pdf

An important issue to note is that there are some deficiencies in the transcripts, especially those by FMP, and hence in the I-CeM data. I-CeM also contains a number of other constraints which limit the accuracy of some spatial and other coding. For England and Wales these constraints are summarised in a second edition of the *I-CeM User Guide*[4] and their effects on the BBCE are outlined in WPs 1, 3, and 4. For England and Wales the BBCE has made significant improvements to the I-CeM data to overcome most of the deficiencies. These improvements give major enhancements to infill I-CeM gaps for entrepreneurs. Other *general* improvements to I-CeM deficiencies were part of a process made in-house at the University of Cambridge, Cambridge Group for the History of Population and Social Structure (Campop) as part of the BBCE project, by the research team on Alice Reid's ESRC-funded project 'An Atlas of Victorian Fertility Decline', and by others in a Campop user group. Important guidance and assistance for these improvements came from the original I-CeM team, Kevin Schürer and Eddy Higgs, PIs of the I-CeM project.[5] These improvements will be part of a version 2 of I-CeM at UKDA planned for 2020. However, for Scotland no general I-CeM database improvements had been made when the BBCE for Scotland was constructed, so that the BBCE construction had to start with the original I-CeM material.

A further point of distinction is the gap of I-CeM for Scotland which does not include the data for the 1911 census year. In order to make comparison with England and Wales, to partially overcome this gap, the BBCE analysis below uses the published census tables for 1911 so that analysis at the aggregate level can be compared with England and Wales; but the 1911 data prepared for Scotland are not part of the BBCE deposit as they are not available at individual level. In contrast, I-CeM Scotland does have the advantage of including the 1871 census data that is missing from I-CeM England and Wales (though 1871 data on entrepreneurs are included in BBCE using other sources than I-CeM).

The I-CeM data for Scotland also differ in having a reduced set of I-CeM spatial codes compared to England and Wales. I-CeM Scotland has codes for census Divisions, Counties and Parishes, but at the time of the construction of BBCE there were no codes for Registration Districts (for which there are no complete maps), nor for Sub-Registration Districts (which did not exist in Scotland). At the time of analysis for this paper there were also missing data for household relationships in the I-CeM data for the variables for RELA

[4] https://www.essex.ac.uk/history/research/icem/documents/icem-guide-version-2-2015.pdf
[5] Schürer, K., Higgs, E., Reid, A.M., Garrett, E.M. (2016) *Integrated Census Microdata V.2 (I-CeM.2).*

and RELAT; these variables that do identify heads are unavailable for some parishes: these include most of Banff and Wigtownshire in 1861.

As in I-CeM England and Wales, the I-CeM Scotland data will be improved in subsequent updated versions, with a major enhancement expected in 2020. However, at the time of analysis reported in this Working Paper I-CeM Scotland data and codes were as first deposited, except for some additional sources made available by Kevin Schürer. However, BBCE contains some improvements that will be included in and subsequent versions of I-CeM for Scotland.

1.3 The BBCE database

The key aim of the BBCE database for Scotland is, as in England and Wales, to provide a cleaned and supplemented database that identifies all employers and own-account business proprietors contained in the census data. All employers and own-account businesses constitute all self-employed individuals, which we define as entrepreneurs (see Bennett et al., 2019a, chapter 2). The BBCE allows other researchers to use the data on entrepreneurs without having to confront many of the problems, and the sources of bias, in the original census data collection, or in I-CeM. However, users need to be aware of the deficiencies in the original data so that appropriate methods are used. The BBCE data largely overcome the gaps from non-available, missing or truncated entries. BBCE also contains cleaned and aligned data, most notably to supplement the early census data for 1851-81, so that a continuous series of entrepreneur information is available for users for 1851 up to 1911 (1901 in Scotland) so that users do not have to embark on complex searchers and re-coding of the original records. The user can therefore go directly to BBCE for files on individual business people that are cleaned, corrected, re-coded in various ways, and supplemented to infill lost, truncated and missing material; these can then have the additional demographic information in I-CeM added for the identified entrepreneurs. The entrepreneur data is given for the RecID of each individual: this allows the entrepreneur information coded into BBCE to be linked to I-CeM. The *BBCE User Guide* gives the full information on the entrepreneur data now available for Scotland, as well as England and Wales. It is not claimed that the final result is a database that can include all businesses, or that it is fully accurate for individuals at the

individual level, but it seeks to come as closely as possible to meeting this target with the information available from the censuses.

A warning is given for users of the supplemented data for 1851-81, where the lack of full responses to the employer and masters' question presents numerous challenges. The final data in the BBCE data deposit listed as the EMPSTATUS_IND variable, and within this the identified status of farmers, must be treated as preliminary. This warning especially applies to small parishes and small sub-categories of the population.

The rest of this paper is divided into four sections. Section 2 discusses the census, I-CeM data sources, and the data challenges that need to be overcome to provide a complete database of entrepreneurs, including data extraction, cleaning, and supplementation for the years 1851-1901. This section also describes how the gaps in FMP and I-CeM information of own-account proprietors for 1891 and 1901 was infilled. Section 3 discusses how the data for 1891 and 1901 are analysed to identify non-response and mis-allocation biases, and how these were corrected. Section 4 describes how the data for 1911 was infilled from the published census reports and assesses the resulting estimates of trends for 1891-1911. Section 5 extends the reconstruction method developed in England and Wales (see WPs 9 and 9.2) to supplement entrepreneur responses 1851-81 in Scotland so that *estimates* of a complete time series of entrepreneur data is available for the whole period. Section 6 reviews the resulting trends. Section 7 records how business partners within the census records were identified and coded (see also WP 18), how company directors for Scotland were identified and data for them enriched using the *Directory of Directors* (see WP 14), and also how portfolio businesses were identified and coded. Section 8 then describes how an urban classification of the I-CeM data was developed for Scotland to parallel that in England and Wales (as reviewed in WP 6).

Downloads of intermediate calculations and the cut-off definitions for farmers are provided as downloads kinked to this paper.

2. The data

The Scotland data were based on I-CeM version 1 and were available for 1851 through to 1901. As noted above, unlike I-CeM England and Wales, from I-CeM Scotland 1871 was available, but 1911 was not available. There were several key data gaps with the Scottish I-CeM that had to be fixed before BBCE data extraction could take place. These were:

(i) Errors in I-CeM spatial coding, reallocation to correct parishes, and removal of duplicates (as in England and Wales: see WP 1),

(ii) Truncation of the occupation string in the censuses 1851-1881 (also as in England and Wales: see WPs 1 and 3),

(ii) Lack of the 1851 occodes in I-CeM Scotland, and

(iv) The omitted transcriptions of own-account codes in the EmployCode variable in the 1891 and 1901 censuses in I-CeM Scotland.

2.1 Improvements to I-CeM

There are several improvements needed to I-CeM spatial coding, reallocation to correct parishes, and removal of duplicates before the data can be analysed. At the time of the database development for BBCE for Scotland only the original I-CeM version 1 was available. Resources were insufficient to make all the improvements necessary, but a number of key enhancements were made.

Checks on the Scottish I-CeM data against published tabulations and population totals by location show deficiencies in I-CeM compared to the Scottish GRO tabulations. Subsequent analysis demonstrated that these included misallocations between parishes, inclusion of duplicate records from FMP that had severe distortionary effects in the parishes with small populations, and a number of missing records (or missing occupational recording). These deficiencies are similar to those deriving from the FMP transcripts and I-CeM parish coding in England and Wales, as noted in the second edition of the I-CeM User Guide.[6] Additionally errors were found in the I-CeM parish dictionary for Scotland. Difficulties for rural Scotland were identified with considerable help and input from Michael Anderson and Corinne

[6] Schürer, K., Higgs, E., Reid, A.M. and Garrett, E.M. (2016), *Integrated Census Microdata, 1851-1911, User Guide version V. 2 (I-CeM.2)*, 2nd. edition, Colchester: Department of History, University of Essex, pp.113-4.

Roughley, continuing their work on mapping Scotland's population history.[7] They developed improved spatial allocations for the development of a detailed analysis of the farming and crofting communities in Scotland. The work by Anderson and Roughley was focussed on the main farming localities, and the modifications to I-CeM they have identified have been included in BBCE. There are doubtless some deficiencies in other areas that have not yet been detected. Identification and correction of all the detailed deficiencies is a substantial task that was not possible during the preparation of the Scottish data for the BBCE data deposit and subsequent analysis. However, the major deficiencies were as far as possible overcome in the data deposit. This included:

- Correction of the I-CeM parish dictionary (thanks to the input from Corinne Roughley: see Roughley (2019a, b));
- Identification of all duplicates detectable, and further checks in farming and crofting areas identified from very detailed checks by Mike Anderson. These include over 16,000 duplicated records in 1851, and over 13,500 duplicated records in 1881;
- Reassignment of individuals to their correct locations where identified (mainly for farming locations). This affected over 25,000 individuals in 1851, 2,000 in 1881, and 6,000 in 1901. This included all individuals, not just farmers in these parishes.

These latter two corrections were applied as fully as possible to 1851, 1881 and 1901, but 1861, 1871 and 1891 were not checked as intensively and will still contain some misallocated data that need future attention.

2.2 String truncations

The occupation strings for the Scottish I-CeM data derived from FMP were truncated at 100 characters in 1851, at 60 characters in 1861 and 1871, and at 80 characters in 1881. This is similar to the England and Wales I-CeM data and derives from the source material from FMP (see WPs 1 and 3). This poses a critical problem for identifying entrepreneurs since employers who reported their workforce often recorded long text strings, meaning that important information (especially on their employees) was lost due to being at the end of each line. In order to recover this information, all strings that measured at the truncation number or at truncation minus 1 (since the final character could be a blank space) were accessed from the original CEBs by research assistants working at The National Records of Scotland, who compared the transcription to the scanned CEBs available from *Scotland's*

[7] Anderson (2018); see also Campbell and Devine (1990); Carter (2001); Devine, T.M. (2018),

People.[8] In addition to filling out the string, they were instructed to correct the original transcription if wrong, to check for split lines, ensure that I-CeM had allocated the transcription to the correct person, and to check sex and age. This effort corrected 7,958 occupation strings across the four early Scottish censuses. Their importance is evident from Table 1, which shows that over 80 per cent of these strings were employers, whose workforce would not have been recorded correctly, without this effort.

Year	Corrected strings	% employers
1851	395	71
1861	2,187	78
1871	5,265	88
1881	111	71
All years	7,958	84

Table 1. Corrected occupational strings in the Scottish census.

2.3 1851 Occode recoding

The original I-CeM deposit did not include the Occode variable for 1851. We are grateful for receiving a version of this variable from Kevin Schürer. However, a much higher proportion of people had the occode 999, the code for unknown occupation. In 1851 653,566 people, or 22.5 per cent of the total I-CeM population, were coded to 999 compared to just 0.14 per cent of the 1861 Scottish I-CeM population. Included in the 999 code were a number of major occupation strings such as 'weaver' (16,478 people), 'servant' (15,425), 'handloom weaver' (14,352) and 'dressmaker' (11,863). It was necessary to recode the majority of these miscoded individuals in order to produce an adequate occupational coding. This was done using three methods. First, every occupation string describing 20 or more people was checked and coded by hand. This recoded 432,485 of those coded to 999. Secondly, the occupational coding of the strings in the other Scottish, and English and Welsh I-CeM census data was used to recode the remaining individuals whose occupation strings exactly matched string in these other censuses. This recoded a further 35,948 individuals. Thirdly, two approaches were used to allow fuzzy matching between the 1851 and 1861 Scottish census occupation strings

[8] Tobias Lund, Annette Mackenzie, and Amber Stevenson.

to identify strings which were similar to those in 1861. The first used wildcard searches on every 1861 string which described more than 100 people to match 1851 999-coded strings to possible 1861 strings; if the 1861 string was found at the start of the 1851 string it was assumed to be the primary occupation and the appropriate occode used. For example, the 1851 string 'accountant bookseller & stationer' matched to the 1861 common strings 'accountant', 'bookseller' and 'stationer': as 'accountant' was the first occupation mentioned it was coded to the 'accountant' occode. This follows the census forms which instructed individuals to state their primary occupation first. The second approach was to use the Stata program *Matchit* (Raffo, 2015) to match the remaining 999 strings to potential 1861 strings on the basis of the similarity between those strings. Hand checking showed that a 'similscore' greater than 0.4 was a valid cut-off for strong matching, and so every string match with above this point was assumed to be correct. These two methods recoded a further 162,728 individuals leaving 22,405 people coded to 999, just 0.77 per cent of the 1851 Scottish I-CeM population.

2.4 Comparing extracted and published data: Re-weighting 1861 farm numbers

In England and Wales a detailed comparison was made of the published census report numbers for entrepreneurs with those contained in I-CeM. This allowed various deficiencies in I-CeM and extraction processes to be identified. The comparisons were particularly important is identifying missing data entry to I-CeM and undertaking infills (see WPs 3 and 13). In Scotland farmers are the only entrepreneurial occupation where the full population is extracted in BBCE for 1851-81 which can be used to compare to the GRO published tables and to act as a check on I-CeM.

There are some difficulties in the comparisons of farmers because there were some inconsistencies in whether or not crofters were included in the published tables under farmers. There is also some doubt about how GRO tabulated the special Scottish categories of tenant, cotter, lotter, pendicler[9], feuer, small tenant, especially in 1861. In 1851, when the Scottish data were tabulated by the Census of Great Britain, it was stated that crofters were included in the farmer category if they returned a specified extent of land, while crofters without acreage or with an inconsiderable amount were included with the agricultural

[9] Cotters, lotters, pendicler and feuer were special forms of land tenancy, and all can be treated like crofters. The spelling of pendicler is sometimes spelled pendicle; the pendicle is the actual piece of land, and the holder a pendicler. Pendicler is the most common spelling used by census respondents in their occupation strings.

labourers. The acres cut-off to be included with farmers was not specified, and their acreage varied considerably: 51 per cent declared no acres and appear therefore to have been definitely excluded by GRO, 10 per cent had 2 acres or less, while 5 per cent had 10 or more acres. From 1861 onwards, the Scottish data were tabulated by the separate Scottish census, which retained similar occupation categories to those used in England and Wales. While there were no specific notes on whether the crofters were included with the farmers, excluding them leads to large deficits in crofting counties specifically, which indicates that the Scottish GRO between 1861 and 1881 included them as farmers instead of as agricultural labourers. However, the exact procedures used by GRO differed between 1861 and the later years. In 1871 and 1881 the total of farmers and crofters extracted from I-CeM was very close to the published total. In 1861 however, about 10 per cent of farmers was missing.

The change in procedures is clear from Figure 1. This shows the total farmers and crofters in I-CeM against the published figures between 1851 and 1881. These count all farmers and crofters, not only the entrepreneurs, as this was the only measure available for all years that compares with GRO. In 1871 and 1881 the farmers and crofters extracted from I-CeM track the published totals closely. In 1851 the GRO appear most likely to have included most crofters with the agricultural labourers as they either did not report any acreage or only a small amount; without these crofters the total farmers in I-CeM matches this number well.

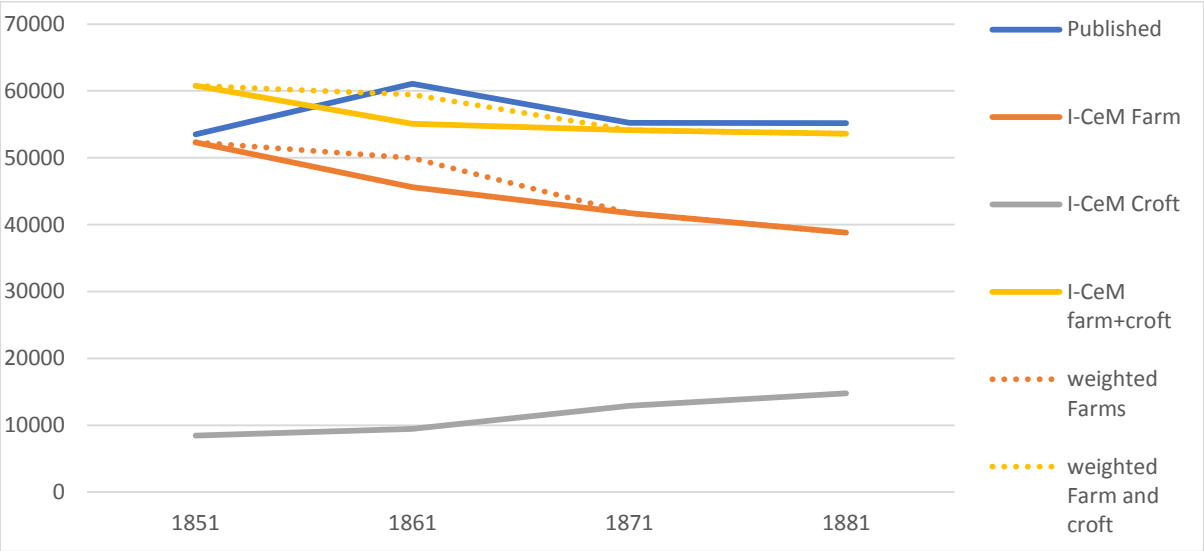


Figure 1. Trends in total farmer numbers, 1851-1881.

However, for 1861 there is a discrepancy based on the different methods that result in this year having an erroneously low estimate of farmer numbers. It is believed that this gap was a result of changed procedures in Scottish census collection for that year rather than I-CeM. It is not clear why these differences arise as there is no mention of changed tabulation procedures in the census reports. It may result from a different treatment of some of the categories of tenant, cotter, lotter, pendicler, feuer, small tenant, as well as crofter. This requires further investigation, although the exact way that GRO operated is unlikely to ever be fully determinable. As a way of dealing with the issue for analysis purposes, weights were created at county level to be applied to the farmers, which bring the total farmers close to the published numbers at county level.[10] The weights are presented in Table 2. Weights for three counties have values below one, where there were more farmers in I-CeM than in the published figures. The use of these weights affects a total of only 95 farmers. They reflect a correction of I-CeM which may not be a GRO issue but one arising from I-CeM misallocations. As shown in Figure 1, when 1861 is weighted the extracted farmers and crofters track the published totals closely. These weights are used in subsequent analysis their use is recommended to other researchers.

Note that the 1861 weights are calculated by comparing I-CeM data to the published numbers for each county, but as noted above, we remain unsure of exactly how these were constructed by GRO. Also as noted earlier, I-CeM is based on Civil County boundaries. The GRO 1861 published report implies the county occupation totals are by Registration County, because most of the tables in that section of the report are by Registration County, but the actual boundaries used are not explicit for the tables compared here. If GRO did in fact use Registration Counties there will be some slight inaccuracies in the weights given here.

County	**Weight**
Aberdeen	1.075152
Argyll	1.076567
Ayr	1.085208
Banff	1.016667
Berwick	1.021318
Bute	1.048474
Caithness	1.051675
Clackmannan	1.021978
Dumbarton	1.197561
Dumfries	1.095642

[10] These should be applied to the farmers only, since the crofters are reconstructed through the supplementation model rather than the farmer model.

Edinburgh	1.241935
Elgin	1.089502
Fife	1.09465
Forfar	1.090956
Haddington	1.048544
Inverness	1.001283
Kincardine	1.082909
Kinross	1.255814
Kirkcudbright	1.126091
Lanark	1.123045
Linlithgow	1.098214
Orkney	1.097665
Peebles	1.052885
Perth	1.126843
Renfrew	1.06237
Ross and Cromarty[1]	1.457266
Roxburgh	1.0592
Stirling	1.025105
Sutherland[1]	1.960358
Wigtown	1.079365
Nairn[1]	*0.8*
Selkirk[1]	*0.921053*
Shetland	*0.995549*

Table 2. Weights for farmers, 1861 by Civil Counties; note counties with values under one italicised. [1]Note: there are significant missing parishes in I-CeM which could be compensated for in a different way in this table in future analysis: Ross and Cromarty has two large parishes missing; Sutherland is missing the large parish of Tongue. Selkirk and Nairn have significant RD and parish boundary problems which mean that they might be better treated by combining with Roxburgh and Inverness/Moray, respectively. We are grateful for Mike Anderson for suggesting these alternative methods of weighting, although they have not been used here.

2.5 Coding 1891 and 1901 for Missing Own-Account Proprietors

As in England and Wales, the later Scottish censuses for 1891-1911 contained a question on employment status, and these data were published, but only for some occupations and for some spatial units. The question on the 1891 and 1901 census household schedules was the same in Scotland as in England and Wales. However, the data on own-account individuals is missing in the 1891 and 1901 Scotland I-CeM data. For unknown reasons individuals in I-

CeM have coded values for this variable (EMPLOYCODE) if they ticked or wrote worker or employer, but not if they stated that they were working on their own account. Therefore, the I-CeM variable in both 1891 and 1901 has only three values: employer, worker and blank. The blanks are a mixture of non-respondents, non-economically active, and the missing own-account proprietors.

This can be examined further by comparing the I-CeM data with the published reports for these years. In both 1891 and 1901 the report tabulated the data on individuals' occupations broken down by their employment status, gender and county.[11] The tables did not cover all occupations, but they nevertheless illustrate that it is only own-account individuals that are missing, rather than a random mix of all employment statuses. Table 3 shows this comparison for 1891 and Table 4 for 1901, giving the number of workers and the number of employers in the report tables and in the I-CeM data.

There are significant differences between the report and I-CeM in a number of counties, especially with regards to workers in Lanark, Forfar and Renfrew in 1891, and in 1901 in Lanark, Forfar and Edinburgh. However, in these cases the differences are all explained by the variation in occupation coding between I-CeM and the census reports combined with the partial tabulation of occupational information in the reports. For example, in Lanark in 1891 a large number of workers are coded in I-CeM to occupational categories which were not included in the census report table that reported employment status. Such individuals must have been coded by the census clerks to different occupational categories, ones that were tabulated. In other words, while the comparison suggests that there are a great deal of missing workers and employers in the I-CeM data, that impression is an artefact of the difficulties of accurately recreating the occupational coding used by the census clerks. Table 5 illustrates this by comparing the total population of each county. The difference is larger for the 1901 data, but in both cases, there are no major differences in populations and what differences exist are all of the same sign, negative, meaning that there are fewer people in I-CeM than in the published report tables. This suggests that the problem is restricted to occupation coding, rather than that individuals are coded to the wrong parish, and consequently that it affects only own-account people whose employment status was not recorded in I-CeM.

[11] *Census of Scotland, 1891, Tenth decennial census of the population of Scotland taken 5th April 1891, with report, Vol. II. Part I., Parliamentary Papers*, CVIII (1893-4), 540-655; *Census of Scotland, 1901, Eleventh decennial census of the population of Scotland taken 31st March, 1901, with report, Vol. III., Parliamentary Papers*, CVIII (1903), 10-307.

	Workers		*Employers*		*Difference*	
County	***Report***	***I-CeM***	***Report***	***I-CeM***	***Workers***	***Employers***
Aberdeen	37392	33982	5019	4825	-3410	-194
Argyll	7007	5907	1294	1336	-1100	42
Ayr	35692	32705	3524	3772	-2987	248
Banff	4286	3789	975	1035	-497	60
Berwick	2619	2377	673	638	-242	-35
Bute	1973	1515	569	595	-458	26
Caithness	2922	2377	652	705	-545	53
Clackmannan	8514	5727	405	471	-2787	66
Dumbarton	22631	19877	1490	1428	-2754	-62
Dumfries	8171	7004	1236	1378	-1167	142
Edinburgh	79365	72607	8165	8456	-6758	291
Elgin	3853	3449	804	801	-404	-3
Fife	32157	28048	3013	3080	-4109	67
Forfar	77303	69100	4572	4525	-8203	-47
Haddington	3254	3046	742	712	-208	-30
Inverness	6644	6046	1095	1106	-598	11
Kincardine	2744	2396	570	535	-348	-35
Kinross	932	792	99	123	-140	24
Kirkcudbright	3696	3301	683	799	-395	116
Lanark	241069	213942	15208	15892	-27127	684
Linlithgow	5321	5463	642	694	142	52
Nairn	666	716	182	187	50	5
Orkney	1874	1388	340	366	-486	26
Peebles	3005	2579	243	272	-426	29
Perth	17854	17016	2365	2399	-838	34
Renfrew	57388	65121	3782	4848	7733	1066
Ross & Cromarty	4179	3514	1154	1104	-665	-50
Roxburgh	9209	8190	1065	1041	-1019	-24
Selkirk	8290	7097	471	512	-1193	41
Shetland	2312	2298	377	350	-14	-27
Stirling	19520	19319	1708	1881	-201	173
Sutherland	987	890	253	223	-97	-30
Wigtown	3252	2400	561	694	-852	133
Scotland	**712463**	**653978**	**63715**	**66783**	**-58485**	**3068**

Table 3. Civil county employer and worker totals, 1891.

	Workers		*Employers*		*Difference*	
County	***Report***	***I-CeM***	***Report***	***I-CeM***	***Workers***	***Employers***
Aberdeen	57901	52838	4768	4720	-5063	-48
Argyll	10903	8884	1214	1248	-2019	34
Ayr	66427	61868	3611	3830	-4559	219
Banff	6836	5370	974	1011	-1466	37
Berwick	4001	3202	487	545	-799	58
Bute	3045	2265	553	553	-780	0
Caithness	4268	3062	501	555	-1206	54
Clackmannan	10327	9776	401	395	-551	-6
Dumbarton	32771	31012	1499	1545	-1759	46
Dumfries	11978	10771	1026	1128	-1207	102
Edinburgh	122067	111563	7373	7945	-10504	572
Elgin	6386	5545	845	986	-841	141
Fife	61126	55892	3194	3195	-5234	1
Forfar	94101	82574	4192	4018	-11527	-174
Haddington	6461	5835	604	666	-626	62
Inverness	10655	9097	1005	1062	-1558	57
Kincardine	5766	5206	615	604	-560	-11
Kinross	1257	1116	104	121	-141	17
Kirkcudbright	5720	5182	559	659	-538	100
Lanark	415436	398548	16504	18099	-16888	1595
Linlithgow	18554	17189	603	723	-1365	120
Nairn	1188	1078	257	155	-110	-102
Orkney	2896	1622	312	297	-1274	-15
Peebles	3759	3441	205	258	-318	53
Perth	23475	21331	2162	2156	-2144	-6
Renfrew	84271	77745	3826	4013	-6526	187
Ross & Cromarty	8630	4836	722	637	-3794	-85
Roxburgh	10516	9828	836	934	-688	98
Selkirk	7558	6596	374	455	-962	81
Shetland	4124	2321	228	258	-1803	30
Stirling	37446	34303	1746	1773	-3143	27
Sutherland	2631	1773	204	242	-858	38
Wigtown	3576	3248	443	502	-328	59
Scotland	**1145686**	**1054917**	**61911**	**65288**	**-90769**	**3377**

Table 4. Civil county employer and worker totals, 1901.

	Report Population		*I-CeM Population*		*Difference*	
County	***1891***	***1901***	***1891***	***1901***	***1891***	***1901***
Aberdeen	280832	304439	281107	303834	275	-605
Argyll	75003	73642	74340	72824	-663	-818
Ayr	226283	254309	225986	254096	-297	-213
Banff	64190	61488	64372	61419	182	-69
Berwick	32406	30824	32367	30631	-39	-193
Bute	18387	18787	18443	18852	56	65
Caithness	37177	33870	37051	33612	-126	-258
Clackmannan	28432	32029	28629	31991	197	-38
Dumbarton	94495	113865	94372	113584	-123	-281
Dumfries	74221	72571	74217	72621	-4	50
Edinburgh	434159	488796	432273	487632	-1886	-1164
Elgin	43453	44800	43197	44746	-256	-54
Fife	187319	218824	186617	216802	-702	-2022
Forfar	277781	284082	277696	273310	-85	-10772
Haddington	37485	38665	37427	38654	-58	-11
Inverness	89317	90104	88923	89870	-394	-234
Kincardine	35647	40923	35632	40896	-15	-27
Kinross	6280	6981	6276	6974	-4	-7
Kirkcudbright	39985	39383	39941	39332	-44	-51
Lanark	1046040	1339327	1041819	1336720	-4221	-2607
Linlithgow	52808	65708	52649	64750	-159	-958
Nairn	10019	9291	10034	9290	15	-1
Orkney	30453	28699	30173	27724	-280	-975
Peebles	14761	15066	14734	15010	-27	-56
Perth	126946	123283	125895	122654	-1051	-629
Renfrew	290798	268980	289981	268423	-817	-557
Ross & Cromarty	77810	76450	77530	66776	-280	-9674
Roxburgh	54194	48804	53477	48809	-717	5
Selkirk	27353	23356	27373	23385	20	29
Shetland	28711	28166	28516	27780	-195	-386
Stirling	125608	142291	124753	141558	-855	-733
Sutherland	21896	21440	21781	21268	-115	-172
Wigtown	36062	32685	35986	32591	-76	-94
Scotland	**4026311**	**4471928**	**4013567**	**4438418**	**-12744**	**-33510**

Table 5. Scottish populations by Civil County, 1891 and 1901.

Having established that it is only own-account individuals who are un-coded in I-CeM and that they are aggregated with blank responses, a method is needed to split own-account proprietors in I-CeM from individuals who gave no answer to the question, i.e. blanks. There

are two separate identification exercises required for this: first, where the occupation categories were tabulated by employment status in the Scottish census reports; second, where the remaining occupation categories have no published data on employment status in the census reports. Each is dealt with using a different method.

2.5.1 Identification where occupation is listed in published tables

The tables in the census reports for 1891 and 1901 include counts of own-account proprietors and blanks for a selection of occupational categories. These can be used to estimate the missing numbers of own-account entrepreneurs by reorganising the I-CeM data to the same occupational categories and then comparing the number of blanks in I-CeM to the number in the census report. The resulting tables give the total of blanks and own-account proprietors in each county for 248 occupational categories (the same categories) in 1891 and 1901. The method followed is then to calculate the percentage of blanks from the published tables and apply it to the lumped-together own account and blanks in I-CeM. The census publications give tabulations of the question at county level, and at Burgh level. Because the counties include the whole of Scotland, the comparisons with I-CeM used the county tabulations. Each county contains a table with the employment status broken down by occupational category and gender. As a result of the previously mentioned differences in occupational coding between I-CeM and the published reports, there are three possible outcomes of the comparisons of I-CeM with each occupational category in each county. For the individuals concerned:

A. There are published data but no data in I-CeM for the occupation/county concerend: 2,806 cases in 1891.

B. There are data in I-CeM but no data in the report: 1,508 cases in 1891.

C. There are data in both the report and in I-CeM: 7,197 cases in 1891.

Cases here refer to the number of individuals in the published table compared to occupational matches in I-CeM. Nothing can be done about Case A because there are no I-CeM data to work with. Thankfully, this affected few people and few occupations; the largest occupational category affected included only 14 people.

For Case B, the absence of published data means that we lack a guide for allocating between own account and blanks; however, once again, the problem is mitigated by the fact that the

categories affected are small, in over 99 per cent of cases the categories amounted to fewer than 4 people.

As noted above, Cases A and B are caused by how I-CeM has coded occupations and how the Census clerks coded them. The individuals affected are almost all likely to be in I-CeM but have different occupation codes from those given by the Census clerks: the individuals in one category in I-CeM were in another in the report. In some cases, these can be identified accurately at the county level (because there are no other people in that category for that county) and then that category's blank and own-account totals can be used (but this will not be accurate to identify at parish level). For example, in 1891 I-CeM in Shetland there were a large number of women in the category 'Knitter (undefined)' but in the published report this category was empty and those women were in 'Knitter (woollen articles)'; in such cases the total for the second category can be assigned to the first category in that county without the need for occupation recoding. In other cases, nothing can be done, but if the individuals are indeed in I-CeM they will be included in any reallocations used for other occupational categories. This means that they will be included in the final own-account allocations, but remain in a different Occode and may have a slightly inaccurate adjustment applied to them to estimate them as own account or blanks. However, this effect will be small on the scale of the own-account population as a whole. *Users should however, be aware that at the parish level in particular there will be errors in the individuals identified.*

In Case C, where there are both report and I-CeM data, for the majority of occupation categories the number of own account plus blanks are almost exactly the same in the report and in I-CeM. This was not necessarily expected, but it is highly reassuring. A small caveat must be added: the report gave data on all individuals aged 10 and over, while the cleaned I-CeM data used in BBCE includes only those aged 15 and over. This produces a minor discrepancy, although the numerical differences are too small to affect subsequent calculations below.

The problems created by the discrepancies between the published tabulations and I-CeM's occupational coding can be dealt with by using a series of different aggregations and adjustments. Where I-CeM and the report totals for an occupational category match, the report percentage is used to split blanks and own-account. This gives the total own account for a specific county/occupation category. Individuals are then estimated using a logit

regression (as discussed below) from data broken down by gender, county, and occupational category. Where the occupational category totals are different, but in the same sub-order (the next aggregation up in the census report occupational classification), then the percentage that were blank can be calculated for that sub-order in that county for that gender. This can be continued up the levels of the occupational classification as needed. Each step involves more aggregation of occupational categories but allows gender and county specificity to be retained. The following variables were thus created (where categories are the combination of occupation/county/gender at each level):

1. Blank_attributed_Occode which is the best guess using the percentage of Blanks for each Sex, Reg Cnty and occupational category. 11,074 categories
2. Blank_attributed_SUBORDER which is the second-best guess using the percentage of Blanks for each Sex, Reg Cnty and SubOrder. 3,334 categories.
3. Blank_attributed_ORDER which is the third-best guess using the percentage of Blanks for each Sex, Reg Cnty and Order. 1,172 categories.
4. Blank_attributed_CLASS which is the fourth-best guess using the percentage of Banks for each Sex, Reg Cnty and Class. 98 categories
5. Blank_attributed_RegCnty which is the fifth-best guess using the percentage of Banks for each Sex and Reg Cnty. 66 categories.

Having established the total of blanks required (the result of multiplying the percentage of blanks by the total own account plus blanks lumped together), the individuals in I-CeM with no employment status code can be assigned between own account and blank using a logit regression. This regresses a dummy with value 1 if the individual is an own-account proprietor and 0 if they are a 'real' blank as follows for each category using the following covariates:

Logit OA Age Sex Marital_status Relationship_to_head Occode (from I-CeM)

This is for category coefficients. We can then predict, using the same coefficients, the probability of an individual being an own-account entrepreneur and the probability of being a blank. Individuals were then sorted according to their probability of being a blank (the inverse of own account), and the number of blanks required in each county and each occupational category based on the published tables allocated on this basis. The process produces the following results shown in Table 6 for the tabulated occupation categories.

New Employ Code / *Original Employ Code*	*Worker*	*Employer*	*Own Account*	*Blank*
Worker	620,419	0	0	0
Employer	0	66,280	0	0
Blank	0	0	108,034	27,216

Table 6. Results of allocation of own-account proprietors, 1891 published occupation categories.

*2.5.2. Identification where occupation is **not** listed in published tables*

Having estimated the best values for blanks and own-account proprietors for the occupational categories listed in the published report tables it is necessary to deal with the categories where responses to the employment status question were *not given* in the published tables. Two separate exercises are required for 1891 and 1901.

(1) Identification of 1891 own account

There are 121 such categories in 1891 including all the professions, farming, and commercial occupations such as merchants. However, 52 of these were occupational categories in which every individual was a worker, such as the general labourer and commercial clerk categories.

For the remaining 60 categories it was necessary to predict the blank and own-account proprietors using a different approach. Since there is no published data available for Scotland, the method adopted was to use as surrogates the England and Wales I-CeM data for 1891 for locations which had comparable occupational structures, and hence hopefully comparable entrepreneurial attributes, to the different regions of Scotland. The Scottish data were split into three regions, based on those used by previous commentators (for example, Anderson, 2018). These regions were then paired with counties in England and Wales with similar economic structures as shown in Tables 7 and 8.

Regions were used in order to use a general level of comparators that did not require highly specific decisions to be made about England and Wales vs. Scottish locations, occupations,

and entrepreneurism. The choice of comparator counties was made after checking for occupational similarities between different possible candidate counties in England and Wales. The final choice was made on the basis of occupational similarity and population scale, to ensure that the English and Welsh counties were of sufficient size to overcome distortions caused by a small number of data points.

Borders	***Central***	***Highlands***
Berwick	Ayr	Aberdeen
Dumfries	Clackmannan	Argyll
Kirkcudbright	Dumbarton	Banff
Peebles[1]	Edinburgh	Bute
Roxburgh[1]	Fife	Caithness
Selkirk[1]	Forfar	Elgin
Wigtown	Haddington	Inverness
	Kinross	Kincardine
	Lanark	Nairn
	Linlithgow	Orkney
	Renfrew	Perth
	Stirling	Ross and Cromarty
		Shetland
		Sutherland

Table 7. Definition of Scottish regions. [1] It is recognised that the England and Wales comparators used as surrogates are open to criticism, but they are offered as a way to solve an estimation problem, by trying to maximise comparability with the need to keep the regions large in terms of total population. They could perhaps be improved by taking account of evolution over time; for example Mike Anderson has pointed out that Peebles, Roxburgh and Selkirk might be more appropriately compared with Yorkshire in later census years; other later-year comparators might also be better for Edinburgh, E. Lothian, Fife, Aberdeen and Kincardine; but this has not been developed further here.

Scottish region	***England Wales comparator***			
Borders	Pembrokeshire	Montgomeryshire	Westmoreland	
Central	Lancashire	Cheshire	Durham	
Highlands	Cardiganshire	Brecknockshire	Radnorshire	Anglesey

Table 8. Scottish regions and the England and Wales comparator counties used.

With these selected comparators the aim was to estimate a model to allocate own-account individuals and blanks using the England and Wales data, and then apply the estimated coefficients to the Scottish data. This assumes that the same relationships hold in Scotland as England and Wales, but this should be achieved reasonably accurately at the level of generalisation chosen and by using appropriate comparator counties. The same logit regression specification was used as for the assignments made above (section 2.3.1), with the England and Wales data as proxy regions generating the coefficients applied to the Scottish data. This produced a probability score for each individual based on their occupation, gender, and location. Individuals were then ranked by probability of being a blank vs. own account. The appropriate number to be assigned blank status was then derived from the proportions in the England and Wales comparator areas, with the remainder coded as own-account proprietors.

The process produces the following results shown in Table 9 for the 60 unpublished occupation categories that contain own account proprietors. This can be compared with Table 6 which shows the extent of reallocating involved.

New Employ Code ***Original Employ Code***	***Worker***	***Employer***	***Own Account***	***Blank***
Worker	794,569	0	0	0
Employer	0	42,127	0	0
Blank	0	0	26,344	27,971

Table 9. Results of allocation of own-account proprietors, 1891 for the unpublished occupation categories.

(2) Identification of 1901 own account

The process was repeated for 1901 with some changes to the method. For the occupation categories broken down by employment status by gender and county in the 1901 census report, the method was the same. It was also the same for those categories given employment status results in the report but which consisted only of workers. For the remainder, however, the 1911 report was used to provide extra information where the county tables contained

more occupation categories were broken down by employment status and gender than 1901: farmers, engineers, artists, merchants and other finance occupations and laundry services were all tabulated for the first time.[12] For occupations not broken down in the 1901 report but which were tabulated in the 1911 report, the 1911 percentages of blanks and own-account proprietors by county were used as targets for the reallocation process rather than the England and Wales comparators used in 1901.[13]

3. Non-response and misallocation bias for 1891 and 1901

The process described above overcame the difficulties of FMP/I-CeM aggregating own account and blanks. In effect the outcome of this effort produced a version of the 1891 and 1901 Scottish I-CeM data which was then equivalent to that available for I-CeM England and Wales. However, as in England and Wales, the responses to the employment question in both 1891 and 1901 contain two difficulties: (i) that non-responses were biased towards certain categories which have to be corrected; and (ii) for 1891 there is a misallocation bias arising from respondents' misunderstanding the question (see WPs 4 and 11; and Bennett et al., 2019b).

To correct for these biases two further steps are required. First, it is necessary to construct data weights to deal with the non-response bias to the employment status question. Secondly, the 1891 misallocation bias has to be corrected because some people erroneously gave their status as 'employer' for a combination of reasons. These are the same problems confronted for England and Wales and they were resolved using the same methods.[14]

3.1 Adjusting for 1891 and 1901 Non-response bias

Non-respondents were not randomly distributed. As in England and Wales non-response was much higher for individuals other than the household 'head' (the person who filled in the census form), such as adult sons, daughters, other relatives, lodgers, etc. (WP 4; Bennett et

[12] The totals for all Scotland are given in *Census of Scotland, 1911, Report on the twelfth decennial census of Scotland, Vol. II, Parliamentary Papers*, LXXX (1913), 264-97; the county data are found in *Census of Scotland, 1911, Report on the twelfth decennial census of Scotland, Vol. I*, CXIX (1912-13).
[13] The occupation categories used in the census tables changed somewhat in between 1901 and 1911, only in cases where the category did not change between 1901 and 1911 could the 1911 breakdown be used.
[14] See Bennett et al., 2019a, b. Further details in WP 4, with adjustment weights given in WP 11.

al., 2019b); this especially affected female relatives who were not heads, and sectors with high female activity, such as domestic service, and many professions. Non-responses reflected a variety of reactions to the question. Some heads assumed that individuals employed in a home-based business were not to be counted, others will have taken their activity for granted or felt it was not 'real' employment. Some heads clearly assumed that the employment status of various household members was obvious and did not need to be stated, most notably for domestic servants. In other cases, a range of other specific instructions was given in the census which may have distracted respondents from responding to the specific and general instructions; for example, professions, commerce and mining. This was further encouraged by the status question being ambiguously defined as it referred to 'trade and industry'. As a result, many individuals probably believed that the question did not apply to them. This seems to have particularly affected the professions, which had especially high non-response rates. Finally, other heads will simply have not answered for reasons unknown to us, or because they missed or ignored the question. There could also be enumerator deficiencies.

Weights were calculated to compensate for the non-responses using a logit regression for the probability of whether an individual was a respondent or not to the employment status question, using a range of independent variables to control for the main features that correlate with the observed non-response bias in the data: gender, relationship with the head of household, and occupational sector. The logit gives the probability of response, providing a means to weight respondents in the same 'non-response class' to compensate for others for those who did not respond. For example, if there is a non-response rate of 0.25 the average weights for that response class should be 1/0.75, and so on. At the individual level, the logit regression calculates the probability of being a respondent. The inverse of these probabilities become the weights used for each observation (see WP 11 for England and Wales). The RESPONSE variable was *one* if the individual gave a response: worker, employer, or own account and *zero* if a true blank was answered. The logit regression for each occupational category (248 in 1891) was run:

logit RESPONSE i.Sex i.RELA_10 for each _1891S or _1901S

An example of the regression results for 1891 for the five occupations with the greatest number of employers (Farmers, Carpenters, Grocers, Tailors and Milliners) is shown in Table 10.

	Farmers	Carpenters	Grocers	Tailors	Milliners
			RESPONSE		
SEX					
Male	0	0	0	0	0
	(.)	(.)	(.)	(.)	(.)
Female	-0.472***	-1.061***	-0.0831	-1.848***	-0.934
	(-16.09)	(-11.19)	(-0.80)	(-7.57)	(-1.20)
RELA. TO THE HEAD					
Head	0	0	0	0	0
	(.)	(.)	(.)	(.)	(.)
CFU member	-1.220***	-4.229***	-2.228***	-1.121***	-2.677***
	(-28.86)	(-10.26)	(-10.18)	(-11.16)	(-5.32)
Older generation	-1.956***	-5.460***	-3.344***	-2.806***	-4.960***
	(-13.03)	(-10.80)	(-8.33)	(-10.01)	(-8.99)
Siblings	-0.896***	-2.584***	-1.596***	-1.732***	-0.633
	(-12.51)	(-4.63)	(-3.87)	(-8.52)	(-0.89)
Other family	-1.193***	-4.652***	-3.216***	-1.649***	-5.654***
	(-6.80)	(-10.40)	(-11.26)	(-7.65)	(-11.23)
Servants	-0.140	-5.997***	-2.885***	-1.243**	-6.887***
	(-0.25)	(-12.74)	(-4.63)	(-2.68)	(-11.92)
Working title	-0.228	-4.155***	-3.128***	-1.201*	-6.136***
	(-0.36)	(-3.80)	(-5.67)	(-2.34)	(-9.32)
Lodgers/boarders	-1.468***	-4.505***	-3.130***	-1.321***	-4.860***
	(-8.47)	(-10.65)	(-14.28)	(-11.00)	(-9.61)
Non-household	-3.055***	-7.819***	-5.150***	-4.348***	-8.069***
	(-12.91)	(-16.59)	(-20.99)	(-27.90)	(-15.27)
Unknown Rela	-2.056***	-5.909***	-4.357***	-3.204***	-6.933***
	(-18.27)	(-13.04)	(-16.97)	(-17.46)	(-13.77)
	1.832***	7.924***	6.111***	4.602***	8.343***
Constant	(135.90)	(19.26)	(29.89)	(59.52)	(9.07)
Observations	56972	25520	26430	30596	51975

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10. Logit estimation of RESPONSE vs. non-response to the employment status question in 1891 for the five occupational categories with the most employers; base categories omitted Male (Sex) and Head (Relationship to the Head).

3.2 1891 Misallocation bias

Misallocation bias for the 1891 census led to a significant over-estimate of the numbers of employers. This has been noted by various commentators at the time, where GRO themselves were critical of the responses received, and in subsequent analyses.[15] However, as Schürer (1991) notes, GRO criticisms were exaggerated because of their dislike for including the question (see also Higgs, 2005). Nevertheless, there are significant issues caused by people misunderstanding or misinterpreting the question posed. In England and Wales this misallocation bias has been analysed at length and a correction method developed. Initially extensive cleaning is employed to remove obvious mis-responses; e.g. domestic servants who responded as employer, or those on living on own means who responded as own account. This cleaning reduced the misallocations very considerably. For the remainder a combination of three alternative methods was used. The preferred method was a robust logit regression model based on 1901 census responses to allocate the 1891 respondents between employer and own account taking account of the most significant explanatory variables for employer status in 1901. The variables used were occode, gender interacted with marital status, population density of the Registration Sub-District, number of domestic servants, and weights based on household relationship codes (RELA). This preferred method was then supplemented for some occupational categories by either simple linear extrapolation of the change in ratio between employers and own account between 1901 and 1911 applied to 1891; or by the average of the ratio between employers and own account for 1901 and 1911.

In Scotland tests of results suggested that only the preferred method was needed: a logit regression that gives the probability of the binary employment status between employer and own account using the Scottish 1901 data, with the calculated coefficients then applied to predict 1891 employers. The model was the same as for England and Wales, except that county rather than RSD was used as the spatial control since RSD is not available in I-CeM:

logit Employers NewOccode Sex Marital status Servts County [pweight=Weights]

NewOccode and County dummies; Sex and Marital status interaction.

[15] *Census of England and Wales, 1891, Vol. IV General Report, with Summary Tables and Appendices*, *Parliamentary Papers*, 1893-4 (CVI), 36.

Table 11 shows the estimates for the 1901 regression that is used for re-allocating the 1891 excess of employers, showing all covariates but only the five most numerous occupations for employers.

NEW OCCODE	
Farmers	4.038*** (9.12)
Grocers	2.109*** (4.76)
Carpenters	2.364*** (5.33)
Tailors	2.188*** (4.93)
Drapers Linen	2.310*** (5.21)
SEX	
Male	0
Female	-1.286*** (-49.31)
MARITAL STATUS	
Single	0
Married	0.500*** (36.07)
Widowed	0.250*** (10.28)
Female # Married	-0.930*** (-20.14)
Female # Widowed	0.374*** (10.07)
SERVANTS	
Servts	0.768*** (60.52)
COUNTIES	
Aberdeen	0
Argyll	-0.212*** (-5.44)
Ayr	0.0650* (2.29)
Banff	-0.260*** (6.71)
Berwick	0.0407 (0.77)
Bute	0.227*** (3.75)
Caithness	-0.226*** (4.40)
Clackmannan	-0.160* (-2.32)
Dumbarton	0.245*** (6.09)
Dumfries	-0.202*** (-5.14)
Edinburgh	0.158*** (6.33)
Elgin	0.197*** (4.10)
Fife	0.0457 (1.54)
Forfar	0.0222 (0.80)
Haddington	0.302*** (5.66)
Inverness	0.0440 (1.20)
Kincardine	-0.0777 (-1.65)
Kinross	-0.0977 (-0.86)
Kirkcudbright	-0.128* (-2.54)
Lanark	0.421*** (18.95)
Linlithgow	0.186*** (3.32)
Nairn	-0.133 (-1.46)
Orkney	-0.553*** (-10.65)
Peebles	0.217**(2.58)
Perth	0.0975** (3.02)
Renfrew	0.245*** (8.47)
Ross and Cromarty	-1.451*** (-35.40)
Roxburgh	0.210*** (4.37)
Selkirk	-0.177** (-2.76)

Shetland	-1.368*** (-19.96)
Stirling	-0.154*** (-3.98)
Sutherland	-1.579***(-21.70)
Wigtown	-0.243*** (-4.79)
Constant	-2.706*** (-6.11)
Observations N	281,229

t statistics in parentheses
* p < 0.05, ** p < 0.01,*** p < 0.001

Table 11. Logit regression of the probability of being an employer over an own account for 1901. After running the regression for 1901, the coefficients are applied to predict 1891 employers. Base categories: Own account, Male (Sex), Single (Marital status), Aberdeen, and civil service (NewOccode).

In order to complete reallocations to correct mis-allocation bias it was necessary to make assumptions to compensate for some missing data. The detailed decisions were as follows:

1. Where data on number of servants in a household was missing, the number was assumed to be zero.

2. Individuals with no marital status code were allocated based on their age: <30 single, 30-70 married, >70 widow/widowed.

3. In 174 cases where the logit gave a perfect probability of one (employer) or zero (own account) these were allocated accordingly.

Although the method was designed to deal with mis-identified employers and reallocate them as own account, sometimes the logit reallocated own-account individuals to employer status. This was expected. This does not occur in England and Wales because of the alternative methods used to the logistic in there. As shown in Table 11, there is a shift not only from employers to own account (36,424) but also a potential shift in a reverse direction from own account to employers (19,524). This corresponds to the degree of uncertainty the model has to confront. For farmers this statistical behaviour would have resulted in an unacceptable decrease in own account from 3,732 to 361 (and a corresponding increase in employers from 29,125 to 32,496). This anomaly for farmers was excluded from the final dataset used in all subsequent analysis (and the reconstructions below). For all other Occodes the shift from own

account to employer was accepted as a realistic outcome. The final reallocation for misallocation bias in Scotland is shown in Table 12.

Reassignment to: / Original from:	Worker	Employer	Own account	Blank	Total
Worker	1,405,664				1,405,664
Employer		71,983	19,524		91,507
Own account		36,424	114,854		
Blank				55,187	55,187
Total	1,405,664	108,407	134,378	55,187	1,703,636

Table 12. Reallocations to compensate for misallocation bias in 1891 between employment statuses using the logit estimator.

These three processes, splitting the own-account and the blanks, weighting for non-response bias, and correcting the mis-allocation bias produces a final dataset for the 1891 and 1901 censuses for Scotland which is consistent with the methods used for England and Wales. The result is a set of weights that are applied to the Scottish data that compensate for the biases and provide estimates of the own account that were not coded by FMP/I-CeM.

4. The 1911 Census

The 1911 Scottish census is not available in I-CeM. However, as noted above, the census report published a breakdown by employment status for many, but not all, occupation categories. This can be used to generate estimates of the total employer, own-account and worker populations for 1911. These estimates can then be used to examine trends in entrepreneurship in Scotland at the aggregate level, even though we do not have the individual level data as for 1891 and 1901. There is, however, the problem that some occupations were not tabulated in the census 1911 publications. To overcome these occupation gaps, the 1901 weighted data (deriving from the compensations outlined above) were used to create estimates of the aggregate totals of employers and own account in 1911.

This is done for aggregations of business sectors/occupations, as outlined in WP 5. That used here is for 17 sectors: EA17. For each EA17 category of aggregated occupations in 1901, the count of employers was subtracted from the sum of the employer after weighting to give the number of blanks accounted for by weighting employers.[16] This number was then expressed as a percentage of the total number of blanks in that EA17 category in 1901; this was then applied to the number of blanks in each EA17 category calculated from the published tables in 1911 and the result then added to the number of employers given in the 1911 tables. The same process was then repeated for own-account proprietors and workers.

This method is necessarily cruder than the approach used for 1891 and 1901 but given the absence of electronic records at the individual level data for 1911 and the incomplete nature of the 1911 census report tabulations there is little alterative available. It is calculated *only for showing 1911 aggregate trends*, and hence to give a comparability with England and Wales. The results are shown in Table 13. This allows tracking of aggregate trends by occupation category. Unfortunately, until the 1911 data are added to I-CeM, there is no way to extend the analysis to the individual level.

Adding the published data for 1911 is valuable for three reasons: as a check that the estimates for 1891-1901 from I-CeM are broadly aligned with the 1911 data; to assess trends over the whole period 1891-1911; and also to permit subsequent comparisons with England and Wales. The trends and comparisons are discussed at the end of this paper.

[16] The 17EA occupation classification is an aggregation of the original I-CeM Occodes for all categories of workers and entrepreneurs: see WP 5.

	1901 % blanks reallocated			*1911 report*				*1911 final*		
	Employer	***OA***	***Worker***	***Employer***	***OA***	***Worker***	***Blanks***	***Employer***	***OA***	***Worker***
Farming & estate work	52.3653	29.6485	23.4677	28511	18645	26934	24306	41239	25851	32638
Mining & quarrying	6.5428	10.9026	156.0079	686	152	143915	14963	1665	1783	167258
Construction	30.0618	21.9136	143.3763	9486	3850	89357	9519	12348	5936	103005
Manufacturing	14.0907	32.7930	413.2192	9378	9789	468702	49482	16350	26016	673171
Maker-dealers	11.0255	42.2328	101.4042	9982	21272	99467	6250	10671	23912	105805
Retail	29.3284	63.3783	105.3866	10350	15076	45684	6915	12378	19459	52971
Transport	6.9053	41.1141	296.0702	2168	2076	63242	15515	3239	8455	109177
Professional & business services	17.4204	26.1790	87.0898	1727	1620	13087	9464	3376	4098	21329
Personal services	2.4144	25.7669	73.8249	2260	5701	18021	67683	3894	23141	67988
Agricultural processing & dealing	10.2129	16.1856	94.2368	1910	754	14199	1245	2037	956	15372
Food sales	14.7368	42.1661	212.4354	13505	14366	78061	8069	14694	17768	95202
Refreshment	3.4283	70.1235	53.3368	4652	3493	10745	15580	5186	14418	19055
Finance & commerce	12.1763	8.6752	79.3110	1711	1715	14254	6106	2454	2245	19097
Public administration, military & clergy			100				55643			55643
Domestic service			100				151734			151734
Undefined general & factory labourers			100				34980			34980

Table 13. Reallocation of blanks in 1911 census report.

Source: BBCE and *Census of Scotland, 1911, Report on the twelfth decennial census of Scotland, Vol. II, Parliamentary Papers*, LXXX (1913), 264-97.

Note: The percentages of blanks allocated to employer, own account and worker status in 1901 are used to distribute the blanks in 1911. See the text for fuller description of the method.

5. Extraction and reconstruction 1851-81

5.1. Extractions 1851-81

Entrepreneurs in the Scottish data were identified, extracted and parsed for the 1851-81 censuses using the same methods as for England and Wales (see WP 3; Bennett and Newton, 2015): a process referred to as 'extraction'. This uses the alphanumeric occupation strings of the whole population to identify candidate employers, masters, business asset owners and other possible entrepreneurs identifiable through their descriptors. It uses a mixture of algorithmic identification and parsing, followed by clerical cleaning and checks of parsing.

The initial algorithm extracting employers with employees extracted 126,401 strings across all years, which was 18 per cent of all unique strings. This was a slightly higher percentage than the original 1851 and 1861 England and Wales extractions, but in line with 1861 England and Wales with completed strings. This is likely to reflect the high quality of the manual transcriptions of the truncated strings in Scotland. Of these, 95,006 strings (75 per cent) were parsed using a second algorithm splitting each employer's occupation from their workforce, while the remainder was parsed by research assistants. The employers' occodes were then checked and re-coded by hand. Additional strings containing acres, masters/mistresses, partners, owners, and directors were extracted from the remainder, cleaned of spurious masters, and their occodes checked and if necessary changed. Farmers and crofters who had not been picked up by either of these extractions (i.e. farmers and crofters without either reported employees or acres) were extracted using their occodes (173 and 174); these were then checked for spurious inclusions. All employers, masters/mistresses, farmers (and crofters), partners and owners of assets other than land were checked and coded for portfolio businesses (see WPs 1 and 3). In addition, strings that had been split over two lines and thus attributed to two separate people were identified and united so that they were assigned to the correct person. Other checks included clerical inspection of the employers with largest workforces against the CEBs. These were checked against the CEBs for non-farmers down to 1000 for males, and down to 20 workers for female employers. Male non-farm employers were checked in the transcripts from 1000 down to 100 employees for odd and suspicious entries. For farmers checks of CEBs were made down to 70 workers for male employers, and down to 20 workers for female employers.

The only key change to this process between Scotland and the earlier England and Wales extractions was the inclusion of crofters in Scotland. This reflects the difficulty of separating crofters and farmers, where the issue of the entrepreneurial status is tackled by the supplementation process (see below).

One feature was peculiar to the Scottish data: some landowners provided the size of their land in merks rather than acres. A merk was an ancient measure of land, corresponding to the amount of land that would cost 1 merk in taxes. As this varied geographically, it is impossible to assign them a number of acres, and these have been left blank. In addition, there were farmers reporting very large estates that were mainly pastoral sheep farming, and included only small parts of arable land. Since the data only provides one measure of acres, these are all counted together, resulting in some huge estates with very small numbers of workers.

In general, the Scottish data extracted on employers from I-CeM have very good fits with the published tables, and hence with the analysis by Rodger (1988) using the published data for Burghs for 1851. However, the Scottish employers data has the same characteristics as that for England and Wales (see WP 13): (i) the GRO published tabulations contain errors and omissions which have not been previously identified by historians, with a tendency for clerks to miss a substantial proportion of the larger employer responses; (ii) the Reports usually had a higher count of the smaller employers, which probably indicates some deficiencies in transcriptions in I-CeM; (iii) however, it is difficult to be sure exactly what the GRO published tables actually show since comparisons using the same definitions stated by GRO fail to produce the same results; while some of this is understandable differences in minor coding and interpretation, for 1851 the differences are considerable; (iv) this indicates that some aspects of the GRO process for collecting and tabulating in 1851 included more employees in the workforce headcount for the smallest firms than the later censuses; this appears to derive from a more thorough effort by GRO in that year to include spouses and other family members that was not fully repeated in other years; however, efforts to re-tabulate the I-CeM information on a similar basis with spouses and working family do not produce fully aligned results for 1851 so that the precise rules followed by GRO are uncertain.

There are two implications of the employer comparisons. First, there is inconsistency in the census processes between 1851 and 1861-81 in the coverage of the very smallest firms, with

1851 including more small employers than later years. This can be overcome by excluding those with the smallest sizes in comparisons (particularly those with one employee, or those with 4 and under). Second, it has to be accepted that apart from any deficiencies in the process by which FMP and I-CeM captured the data from the CEBs, the census process itself had failures in capturing some of the largest employers. The clerks, when confronted with the complexity and scale of the tabulation challenge with the employer data, appear to have not been fully effective in extracting it for GRO publications so that the published tables under-count the largest firms that are actually in the CEB data. It is possible that some of the largest acreages and employers were excluded from the farming tables as they were coded to a status based on rank for the landowners, and hence coded to landowning rather than farming occupation. However, the lack of full documentation of the clerks' procedures mean that it is difficult to judge whether this occurred, though it seems unlikely given the way the census clerks tabulated each published table in separate sweeps.

These issues are separate from the need to weight 1861 farming and crofting as discussed above (Section 2.4).

A point to note is that in Scotland, as in England and Wales, women who were recorded as 'xxx wife', such as 'farmer's wife', 'carpenter's wife' etc., were not included in BBCE. They are excluded for two reasons; first, because this record is ambiguous and it is difficult to control for which are real records: some are economically active, others are recorded just as a status. Second, because the census method of recording 'xxx wife' was very inconsistent in different years and between enumerators, including them leads to significant swings in female participation that are entirely spurious. Further research could be undertaken on these 'xxx wife', although the results are likely to be uncertain. For the BBCE in Scotland, as in England and Wales, the result of excluding them means that the number of female entrepreneurs and the rate of entrepreneurship are slightly under-estimated. This should be borne in mind in interpretations by users. However, the level of female entrepreneurial participation found in census data is nevertheless higher than in almost all other major sources.

5.2. Estimates of farm entrepreneurs: Farm model 1851-1881

The extracted data from the previous methods give only the initial identification of all entrepreneurs. An extensive exercise then needs to be undertaken re-code the information so that all entrepreneurs are correctly estimated with their employment status: as employers and own account, and correctly separated from workers. For farmers this can be undertaken directly from the full information provided by farmers on their employees and acres. For non-farmers a different supplementation process is required.

The extraction process uses various criteria to identify who was or was not an entrepreneur from their occupational descriptors (see WP 1). In the reconstruction process applied to these data the extracted entrepreneurs are coded to different *Extraction Groups*. For farmers the extraction Groups and their assignment between employers and own account is based on that used for England and Wales, discussed in WP 9, WP 9.2, and Montebruno et al., (2019). These Groups are as follows:

Group 1: 'farmer' with stated employees.
Group 2: 'farmers' stating 'emp' but with no employees stated
Group 4: 'farmer' not stating 'emp' or acres
Group 5: 'farmer' giving acres but with no stated employees

The assignment of these Groups between farm employers and own account, as in England and Wales, is as follows:

Group 1 farmers are accepted as employers, but they will be incomplete to the extent that there is any census non-response rate, or database mis-identification from I-CeM because of transcription gaps or truncations; hence, Group 1 will not include all farm employers and needs supplementing from the other groups.

Group 2 is assumed to be own account; although a somewhat ambiguous category their numbers are very small.

Group 4 state no other information than that they are 'farmers'. This is an occupational descriptor. Without other information, that they employed workers or had an acreage, they have been assumed to be farmers working on other farms; hence, all are assigned to worker

status. This is line with how the GRO in London treated Scottish farmers and crofters in 1851 for tabulations.[17] However, those with a portfolio of other activities are assigned to own-account status because it is assumed that their other occupation was in the entrepreneurial occupational categories then they were a small business trader, and the farm statement was adjoint with their other occupation - as a joint own account trader-farmer.[18] The numbers identified in this way were small.

Group 5 contains both employers and own account and an assignment between the two is necessary, as well as separating any workers. The category has the valuable additional information of a specific acreage. Those that had acreage of less than 2 acres and stated no employees were assigned as workers. This assignment follows decisions made when *Agricultural Statistics* were collected, which assumed active farmers had to have 2 acres and over, with those below being 'smallholdings', although definitions varied over time. It is assumed that farms below 2 acres were not viable as businesses, but instead were essentially subsistence farmers whilst also operating as farm labourers elsewhere, or as paid workers in other occupations (such as estate work).[19] This removes those who could not normally have been own-account or employer farmers as individuals.[20]

The rest of Group 5 with 2 or more acres are assumed to be entrepreneurs and the acreages are used to separate own-account farmers from employers who had neglected to return their workforce numbers. The employers estimated from Group 5 were added to the employers from Groups 1 to give all farmer employers; the rest of Group 5 were assigned as own account. This results in all farmers being estimated directly from their census descriptions.

[17] It also largely accords with how they responded in to the status question in 1891-1901, although this can be investigated further for specific locations.

[18] Note that the extraction process for 1851-81 identifies portfolios only if individuals are already extracted as master or employer (Groups 1-6); hence the equivalent of a farmer-worker with a trading portfolio, say a butcher, would not have been picked up as a portfolio if they did not mention being a master or had employees. For the later censuses 1891-1911 a wider definition of portfolios is used: see Bennett et al., (2019b, chapter 11). For the special Scottish cases of farmers, crofters etc., this can be investigated further, especially for those crofters who also were farmers.

[19]This could be investigated further through comparisons with the responses to the status question in 1891-1901, especially for specific locations.

[20] This is an assumption across all the individuals concerned; in a given household there could have been other labour inputs form spouse and/or family; and there could have been other businesses of the household unit in the croft which have not been picked up by the portfolio coding. In England and Wales this was tested by seeing how far adding co-occupants of the household with the same occupational farming description might indicate an employer status of family members that was otherwise not declared; it did not provide any adequate way of coding Group 5 there that fitted with other comparator data (such as *Agricultural Census* or census aggregate farm numbers: see Montebruno et al., 2019), but further analyses for the rather special case of Scottish crofts is certainly merited and is being undertaken by Anderson and Roughley.

It should be noted that all assignments are based on the occupational statement being one of *'regular' or 'main' employment; i.e. it is assumed that seasonal and very limited work is excluded.* This is in line with the date of the census which was chosen to minimise the effect of seasonal work. However, the concept of full- and part-time was undeveloped in this period and for crofting communities would have had difficult interpretations, so that the allocative decisions made here, and their assumptions, can certainly be investigated further and probably improved upon.

The method used to estimate the employers within Group 5 follows the same steps as in England and Wales (WP 9, section 3.5; Montebruno et al., 2019b), but with a few adaptations. A model is estimated of the relationship between the probability of being an employer and the acres declared by Group 5 farmers. The model uses individual farmers who reported employees and acres (Group 1) as an estimator, to allocate farmers that reported only acres (Group 5) using the acreage which both groups reported and a series of other covariates. The model uses a logit regression to assign between employer and own account status, with a cut-off of 0.25 for the logit, as in England and Wales. This allocates Group 5 that contain a proportion of non-responses on employment that should have been reported as employers, leaving the rest as own account. In England and Wales, the model was estimated for each Registration District (RDs) separately to take account of differences between the employer/own account ratios for a given acreage within different parts of the country. However, RD boundaries are not available in Scotland, so parishes were used. This has the advantage of fine-grained detail, but the disadvantage that the number of individuals in each unit is smaller, very small in some parishes, and hence the model is more unstable and difficult to estimate for some locations; the parish level is also more subject to any errors in transcripts of I-CeM coding.

The model reallocates only small numbers from Group 5 into employer status: 7,094 (18.3 per cent) for 1881, 6,943 (16.6 per cent) in 1871, 7,067 (15.49 per cent) in 1861, and 8,658 (16.55 per cent) 1851. For Scotland, the model produces very different results from England and Wales, with many parishes having no reassignment of Group 5. In these parishes, all farmers were Group 1 farmers who gave their employee numbers. In other parishes the contrast is partly due to the small size of many of the parishes, which gives greater variation which tends to lead to higher cut-offs of the level of acres.

As in England and Wales, others engaged in farming as a secondary activity were not included with farmers in this analysis or coded as farmers in the BBCE data; they were coded instead to their main status, which is more accurate. This is different from I-CeM occupational coding where everyone who mentions farming is coded to farming even if it was a subsidiary activity.

BBCE also *excluded* as farmers all those giving their 'farm' occupation as tenant, cotter, lotter, pendicler, feuer, or small tenant,[21] which are special terms used in Scotland. For 1851-81 these individuals were part of the reconstruction process used for non-farmer census supplementation (see Section 5.3); for 1891-1901 they were allocated using the employment status coding they declared.

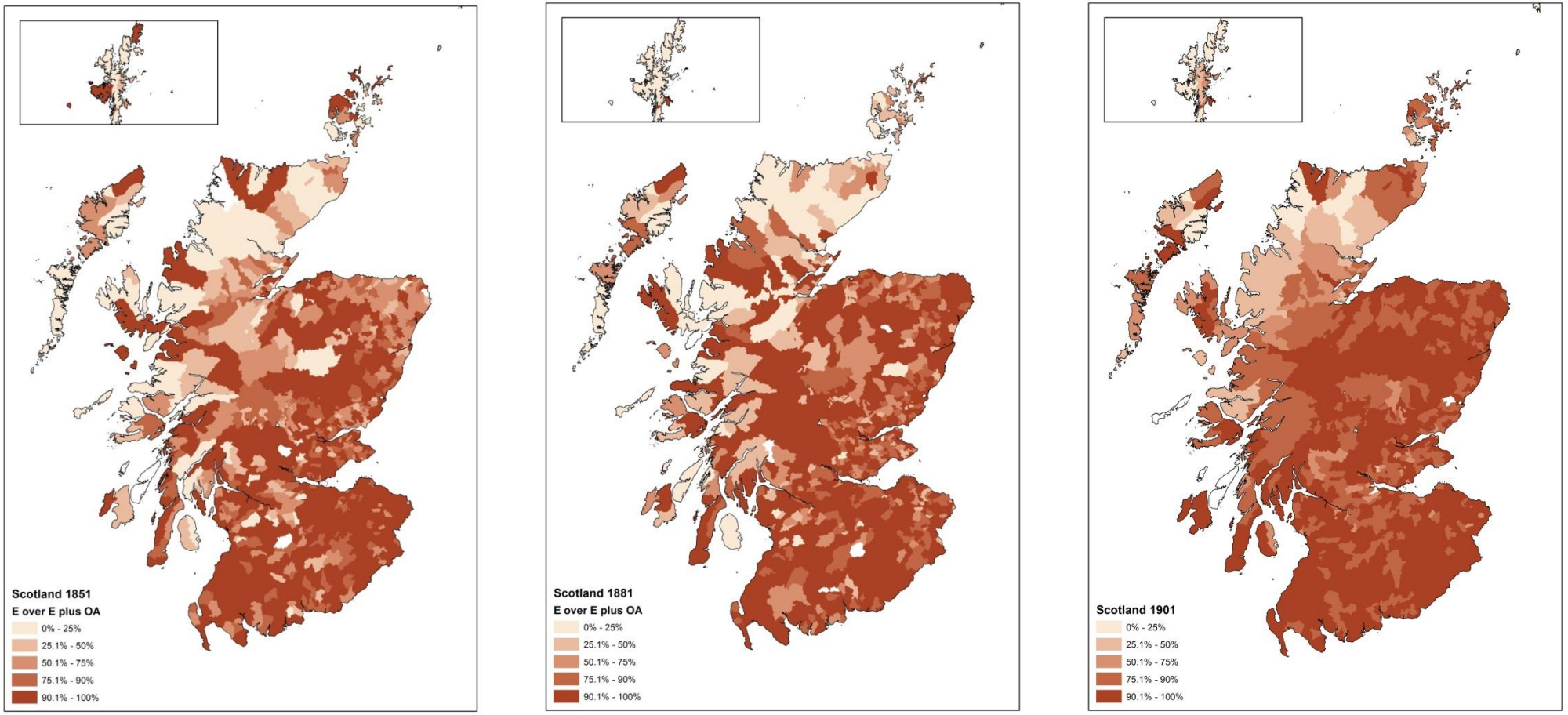
Crofters were also *excluded* (mostly) from the farm model, and were also dealt with using the general supplementation model described below. This decision was made because only a very small proportion of crofts (and almost no tenant, cotter, lotter, pendicler, feuer, or small tenant) gave employees. Hence they are not identified as members of the groups 1, 2, 4 and 5 above. If they had been included, over half of crofters and all of the other categories would have fallen under Group 4 and been allocated as workers, which would have been incorrect. Instead crofters were treated in two ways. (i) Those who gave employees or large acres were treated as farmers and allocated by the farm model. (ii) Crofters with no employees or acres stated were allocated by the general reconstruction and supplementation process. This mainly split them between worker and own-account status – which accords with most analyses of Scottish crofters (Anderson, 2018; Devine, 2018; Carter, 2001). However, it is clear from discussions with Anderson and Roughley, and the results of their ongoing analysis at the time of finalising this working paper, that crofters, and the special Scottish cases of tenant, cotter, lotter, pendicler, feuer, and small tenant were in many cases most appropriately considered as mostly only own account with a smaller, but not insubstantial, number who are likely to be employers (though not stating employees). The farm model used here, therefore, should be treated as the outcome of using equivalent methods to those used in England and Wales. Further developments of the data and its interpretation by Anderson and Roughley will provide a better and more precise specification of these Scottish special cases of crofters, tenants, etc., for those parishes in which they occur. For the data now available in the BBCE

[21] Many small tenants are recorded as 'S.T.' in the CEBs and are coded 999 in I-CeM (as unknown). Hence they were completely excluded in the reconstruction process.

data deposit, however, the estimates developed here provide preliminary estimates of farmers' entrepreneurial status which can be used, with caution applied especially to small parishes in the main crofting areas. They also provide robust estimates of the employer farmers, since adding large numbers of own account with no acreage or other information does not affect how the logit cut-offs are estimated. As a binary model, an employer is only estimated if they had certain characteristics, which from the data provided in the census transcripts, which none of the excluded crofters, tenants and others possessed.

The results of the identification of farm entrepreneurs are shown for the three years 1851, 1881, 1901, in Figure 2. This is based on the updated and corrected I-CeM data at parish level constructed with the help of Anderson and Roughley, but not their improvements to the crofting, tenant and other special cases. As well as the preliminary nature of these farm data, there remain other limitations because there are missing data in some parishes, and the numbers are so small in some parishes that the results can become extreme and unstable. The large size of many parishes, especially in the Highlands, also can be misleading as it averages a pattern over a wide area that may only reflect a small part of the area that was actually farmed. Nevertheless the figure demonstrates that an overview of Scottish farm developments can be obtained from the farm model. The maps show the employers in each parish as a percentage of all entrepreneurs (employers plus own account) after reconstruction.

Figure 2. Employers as a percentage of all farm entrepreneurs (employers and own account), by parish 1851, 1881 and 1901.



In 1851 the traditional concentration of higher levels of employer farming with hired labour in the Central Valley and Southern Uplands is evident, as well as in many of the better lands around the edges of the Highlands mountains and seaboard, where many areas had 75 per cent or more employers among farm proprietors. However, the area where employers were dominant over own account was fairly restricted compared to the following years. By 1881 the area where employers formed the majority had expanded, with even more parishes having over 90 per cent employer proportions. In contrast to developments in England over the same time, although there was a similar slowdown in the number of farm proprietors 1851-81, there was an increase in those employing labourers. This reflects growing demand from urban Scotland and shipments to England, especially London, which stimulated a shift into cattle rearing, and into dairying and pasture close to the Scottish cities and towns. By 1901 this pattern of Scottish farming had consolidated, with more areas in the non-Highland zones intensifying towards employer dominance, whilst in the Highlands fewer areas had over 75 per cent employer farmers, with the vast majority having less than 50 per cent employer proportions.

These maps provide some useful indicators for further analysis. However, their preliminary nature is at best indicative pending improved allocation of the crofters, tenants, cotters, lotters, pendicler, feuer, and small tenants. The most likely updates to this analysis will lead to a general decrease in the estimates of the percentage constituted by farm employers because of improved estimation of the rest of the agricultural population (i.e. increasing the size of the denominator in the calculation of the percentage). This will mainly increase the scale of the own-account farm population, and hence reduce the proportion (though not the number) of employer farmers. Figure 2 therefore probably over-estimates the employer proportions in Highlands and other crofting areas. The new estimates are unlikely to change significantly the number of employers estimated because of the nature of the logit estimation process which requires them to exceed a specific set of criteria that none of the crofting and other special cases meet; only the proportion or employers may change as a result of an increased estimate of the total population of farm entrepreneurs.[22]

[22] We are very grateful for discussions with Mike Anderson and Corinne Roughley on the size of the total own-account population, and for seeing their early estimates of the effects of detailed investigation of the crofter, tenant, cotter, lotter, pendicler, feuer, and small tenant population, which generally confirm the patterns in Figure 2, but differs in detail for some parishes.

5.3. Non-farm supplementation 1851-1881

As in England and Wales the earlier censuses require supplementation to account for those employers and own-account proprietors who did not report themselves as 'employers' or 'masters' because the census question was restricted in its design and scope: this process we term 'reconstruction' (WP 9 and WP 9.2). Two methods to undertake this reconstruction were developed for England and Wales; referred to in the BBCE User Guide as EMPSTATUS_IND (which gives individual level census supplementation), and EMPSTATUS_NUM (which gives aggregate level census supplementation). The former is the most accurate way of estimating the most probable status of actual individuals as entrepreneurs or not (see WP 9.2); the latter is the most accurate way of estimating the most probable population total as an aggregate (with poorer quality at the individual level) (see WP 9). The former uses a tailored logit cut-off, the latter an intelligence-led approach where secondary sources as well as logit estimates are used. WP 19 provides downloads and a guide to how these different supplementation estimates were calculated for both Scotland, and for England and Wales. In Scotland the tailored logit cut-off method developed for England Wales was used exclusively (WP 9.2), with no attempt to use the alternative intelligence-led approach. As with England and Wales 83 additional sub-occodes were created to better separate individuals within complex occupation categories with high variance between employer, own-account and worker status. The supplementation estimates were reconstructed for all census years 1851-1891. The decision process for these cut-offs is given as a download with this WP.

It should be noted that the supplementation method seeks to overcome the imprecision and defects of the census questions of the period. It cannot produce perfect answers to questions that were not fully asked. Users of the supplemented data for 1851-81 in the BBCE data deposit listed as EMPSTATUS_IND must therefore treat these statuses as no more than estimates, especially for small parishes and small sub-categories of the population. Also as noted above, within the supplementation data for EMPSTATUS_IND, the identified status of farmers must be treated as preliminary for the Highlands and Islands.

The reconstruction process based on tailored logit cut-offs uses a logit regression with the 1891 data on employment status to distinguish entrepreneurs from workers. This logit used the weighted 1891 data on status and the variables age, gender, marital status, relationship to

head of household, sub-occode, county population density and number of servants. The coefficients generated by this regression to identify entrepreneurs based on 1891 were then swapped with the 1851-81 data to generate the probability of each economically active individual in the 1851-81 censuses being an entrepreneur or a worker. This was applied to all sub-occodes *apart from farming, but including crofting*, as discussed above (Section 5.2). These probabilities ranged between 0 and 1, where 0 meant no chance of being an entrepreneur (such as a domestic servant or textile worker) and 1 meant someone was definitely an entrepreneur. In order to generate aggregate estimates of the supplemented entrepreneur population it was necessary to choose a cut-off for each sub-occode, namely the probability above which an individual was likely to have been an entrepreneur rather than a worker. While applying the coefficients for 1891 to the 1851-81 data seems at first to use an assumption that the demographic and other covariates were constant, this is adjusted in the next stage by applying variable cut-offs.

The tailored logit cut-off method for reconstruction makes a choice of cut-off tailored to each sub-occode, because the characteristics of entrepreneurs vary by sector; thus, an entrepreneur in dressmaking was much more likely to have a lower cut-off if female (which were often entrepreneurs on own account) than one in coal mining and male (who were mostly workers). For each sub-occode the cut-off to be used determined the total number of entrepreneurs estimated. The most appropriate cut-off was determined by one of two methods. The first took the cut-off which best predicted the 1891 actual numbers in each sub-occode when the 1891 model was used to predict 1891 entrepreneur numbers. The second used the cut-off which gave the total closest to the number of entrepreneurs in that sub-occode to that which have occurred if the 1891 ratio between entrepreneurs and workers was maintained in 1851-81. In both cases the choices were guided by the 1891-01 data: the choice between each comparator was determined by whether the trend produced was sensible; but it was also still intelligence-led in the sense that other secondary data were used in these choices (as in England and Wales: WP 9) in conjunction with the tailoring of the cut-offs choices (WP 9.2). The selection decisions can be inspected in downloads attached to WP 9.2. This also allows some account to be taken of changing demographic and other characteristics that would otherwise be assumed to be held constant from 1891. The cut-off choices made varied between 0.1 and 0.8 reflecting the fact that in some sub-occodes the probability of being an entrepreneur is much rarer. Choice of a high cut-offs also reflects a judgement to identify only entrepreneurs where the level of certainty is high. This seeks to avoid false positives.

This leads to a database that has a higher confidence of accuracy that true entrepreneurs are estimated but will result in some individuals not being identified as entrepreneurs who should be (false negatives). This results in the number of entrepreneurs estimated in the Scottish data being lower in total than it perhaps should have been; as recognised for England and Wales, the tailored cut-off method is good at identifying individuals, but less satisfactory at identifying the total numbers. However, in England and Wales the difference between the two methods are small for employers, though larger for own account where the uncertainties are higher (see Bennett et al., 2020). Table 14 shows the cut-offs used for entrepreneurs in 1851-1881.

Logit cut-off	**Number**				**%**			
	1851	**1861**	**1871**	**1881**	**1851**	**1861**	**1871**	**1881**
0.1	266	254	264	264	33.0	31.4	32.4	32.4
0.15	65	61	65	69	8.1	7.5	8.0	8.5
0.2	53	49	51	55	6.6	6.1	6.3	6.7
0.25	55	56	56	50	6.8	6.9	6.9	6.1
0.3	48	51	49	48	5.9	6.3	6.0	5.9
0.35	52	50	51	54	6.4	6.2	6.3	6.6
0.4	33	35	32	33	4.1	4.3	3.9	4.0
0.45	14	17	17	18	1.7	2.1	2.1	2.2
0.5	63	63	63	66	7.8	7.8	7.7	8.1
0.6	58	67	65	55	7.2	8.3	8.0	6.7
0.7	51	51	51	50	6.3	6.3	6.3	6.1
0.8	49	55	50	54	6.1	6.8	6.1	6.6

Table 14. Cut-offs of 1891 logit estimates for entrepreneurs in Scotland that best fit 1851-81: number for each Sub-Occode and percentage of all 840 Sub-Occodes containing entries.

These cut-offs are similar to those generated for England and Wales and given in WP 9.2 (Table 1). The large proportion of sub-occodes with low cut-offs in both Scotland and England and Wales, with over half having cut-offs below 0.25, reflects the fact that the characteristics of entrepreneurs are quite widely spread in many occupations. It is also a feature deriving from the strong skew of the distribution: in all cases the entrepreneurial

probability distribution approximates a log normal, having large proportions with very low probabilities, and rapidly declining proportions as probabilities increase. The small difference between 0.1 and 0.15 can mean a large change in the number of entrepreneurs estimated, especially in large categories such as dressmakers, grocers or innkeepers.

Having generated aggregate totals of entrepreneurs, individuals were then identified to fill those sub-occodes up to the total number estimated. Extracted entrepreneurs were always coded as entrepreneurs; these individuals were in four groups:

Group 1: anyone with stated employees,

Group 2: anyone stating 'emp' but with no employees stated,

Group 3: anyone described as 'master', cleaned for spurious masters,

Group 6: owners or proprietors of business assets: e.g. mine and quarry owners, ship-owners, but not land/housing

Any additional individuals required were taken from individuals who were not extracted entrepreneurs but whose logit probability was above the cut-off for that sub-occode. The result is a supplemented estimate of all entrepreneurs.

The entire process was then repeated to identify employers: first, a 1891 logit was used to calculate the probability of 1851-81 supplemented entrepreneurs being employers; second, tailored cut-offs were chosen; and third, individuals were assigned to employer status, again extracted individuals were assigned first, in this case only group 1, with additional numbers being made up from individuals not in group 1 but with a probability higher than the appropriate cut off. All remaining entrepreneurs were then coded as own-account proprietors. Table 15 gives the cut-offs for employers produced by this process.

These Scottish employers cut-offs are similar to those for all entrepreneurs, in contrast to the results for England and Wales, reported in WP 9.2 (Table 2), where employers tended to have higher cut-offs. It is likely that the necessity of dealing with the own-account issue in 1891, as detailed above, has rendered the distinction between employers and own-account individuals in the 1891 data less stark and thus it is harder to distinguish between them using a logit model, as a similar model is used in both processes. However, in both Table 14 and Table 15 it is notable that the distribution of cut-offs is very stable across all years, something we would expect given the relative stability of the structure of many sectors in this period.

Logit cut-off	Number				%			
	1851	1861	1871	1881	1851	1861	1871	1881
0.1	427	477	481	491	73.7	75.5	75.5	75.9
0.15	12	12	12	12	2.1	1.9	1.9	1.9
0.2	12	12	12	12	2.1	1.9	1.9	1.9
0.25	3	3	3	3	0.5	0.5	0.5	0.5
0.3	8	8	8	8	1.4	1.3	1.3	1.2
0.35	15	15	15	15	2.6	2.4	2.4	2.3
0.4	19	19	19	19	3.3	3.0	3.0	2.9
0.45	53	55	56	56	9.2	8.7	8.8	8.7
0.5	8	9	9	9	1.4	1.4	1.4	1.4
0.6	11	11	11	11	1.9	1.7	1.7	1.7
0.7	8	8	8	8	1.4	1.3	1.3	1.2
0.8	3	3	3	3	0.5	0.5	0.5	0.5

Table 15. Cut-offs of 1891 logit estimates in Scotland for employers that best fit 1851-81: number for each Sub-Occode and percentage of all Sub-Occodes containing entries.

6. Trends, 1851-1911.

Having identified the missing own account, corrected for non-response and misallocation biases in 1891 and 1901, supplemented the extracted entrepreneurs through the reconstruction process in 1851-81, up-weighted the 1861 farmers in I-CeM, and estimated the numbers of entrepreneurs in 1911 from the published data, it is possible to generate for the first time a long-run time series of the number of entrepreneurs, employers and own-account proprietors for Scotland.[23] Figure 3 show this for all entrepreneurs, employers, own account and workers.

The limitations of the data as a whole, and the supplementation method, should be borne in mind in the interpretation of this Figure. Nevertheless at an aggregate level the data can be treated as a good approximate guide. The trends in absolute numbers are similar to those in England and Wales (see WP 4, WP 9, WP 9.2, and Bennett et al., 2019a, b, 2020):

[23] For previous discussion of the size of the business population see Moss and Hume (1983) and Rodger (1985).

entrepreneur, employer and own-account numbers fairly steadily increased over the period, peaking in 1901; after 1901 employer numbers continued to increase, but there was a sharp fall in total entrepreneurs resulting from a drop in own-account proprietors This decline may be slightly exaggerated by the method of allocating the blanks in the 1911 published data, which uses the simple ratios: between employers, own account, and workers. Our analysis of England and Wales shows that own-account individuals were more likely to give a non-response for employment status; thus, using the simple ratio from the published tables is likely to under-estimate the total number of own-account and over-estimate the total number of employers. However, this effect will be small so that, given the large scale of the changes in own account 1901-1911, if there is an exaggeration it is only amplifying a real trend rather than giving a misleading indication of the general trends.

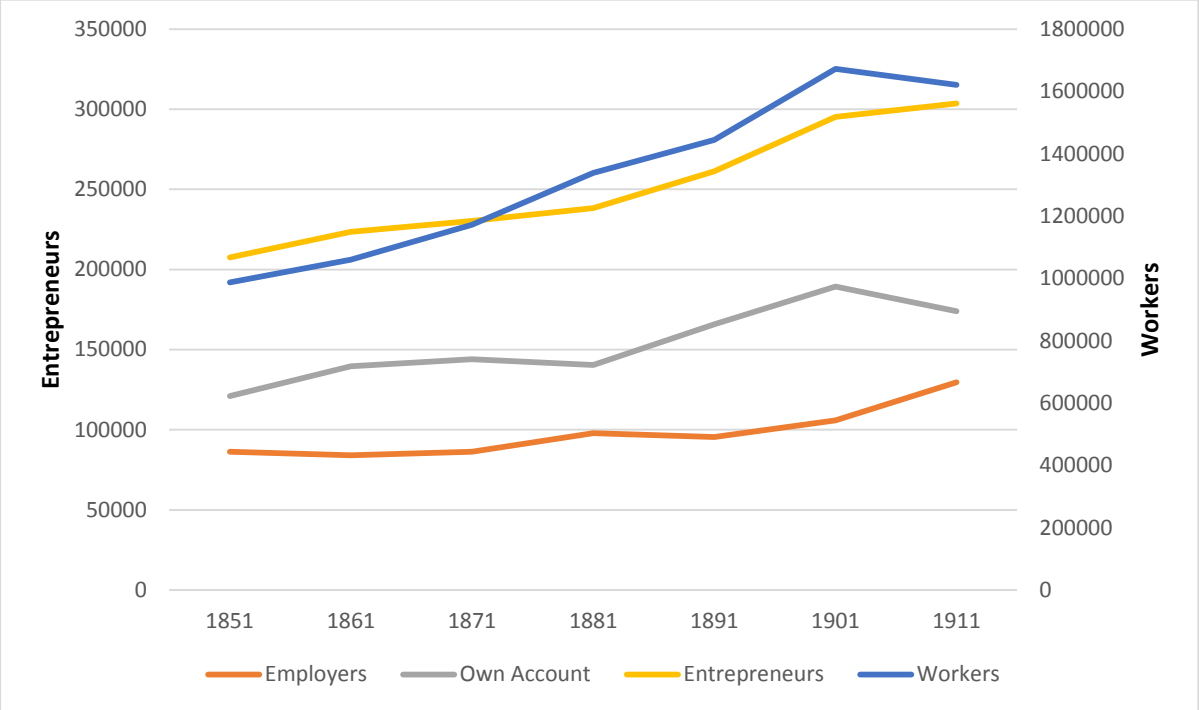


Figure 3. Trends in entrepreneurship in Scotland, 1851-1911, weighted for the small effect of missing 1861 farming data (this is not affected by missing RELAT codes for heads in the 1861 I-CeM data).

The trends in the entrepreneurship rates are shown in Figure 4, calculated by dividing the total in each category by the economically active. The trends here are less marked, than those for aggregate numbers, but show, as in England and Wales, that the three rates declined over

1891-1901, with a more sudden fall 1901-1911 for own-account entrepreneurship accompanied by a rise in the employer rate which drove a slight increase in the overall entrepreneurship rate. Throughout the period the entrepreneurship rate in Scotland was about 2 per cent higher than that in England and Wales, but rose to nearly 4 per cent higher in 1911. It is also an underestimate given that revisions of participation as own account and employers in the crofting and other agricultural tenancies (by Anderson and Roughley) are likely to lead increased estimates of farm/croft entrepreneur numbers; and also because 'xxx wife' are excluded, some of which will be genuine entrepreneurs. The higher Scottish rate is likely to reflect the generally weaker position of labour compared to capital in Scotland which usually resulted in poorer lower wages than in England and Wales; although in some sectors the reverse was sometimes true – e.g. in coal mining where migration to Scotland occurred. Many people reacted to this set of circumstances by emigrating, but an alternative for those in a disadvantageous position of waged labour may well have been to start their own business, resulting in higher rates of entrepreneurship. It also reflected the continued existence of many more isolated communities which required a higher base line of businesses and offered few alternatives for entrepreneurship. This was particularly marked in isolated crofting and fishing communities in the Highlands and Islands (as also observed by Anderson 2018). In England and Wales in contrast, the increasingly dense development of the transport network meant that many settlements could no longer support a full set of small local businesses (shoemakers, tailors, blacksmiths etc.); such goods and services could be purchased elsewhere and transported to a range of increasingly remote locations, which depressed the demand for small local businesses and lowered the entrepreneurship rate. However, in Scotland, transport networks were less well developed and the terrain more challenging and thus settlements throughout much of Scotland required a wider range of businesses and the entrepreneurship thus rate remained higher.

The declining trend of entrepreneurship in Scotland was driven, in part, by similar factors to those which determined the course of events in England and Wales, namely technological developments, mechanisation and business concentration, all of which made it harder to run small businesses, especially own-account enterprises, driving more people into waged labour whilst at the same time more secure and higher waged jobs expanded. Furthermore, even if wages and working conditions were worse in Scotland, they nevertheless improved over this period, making waged labour more attractive. However, in Scotland it was also the case that population growth was concentrated in those locations where waged labour was most

abundant: notably the central valley, Lothian and Strathclyde (Anderson, 2018), and consequently population growth tended to increase the worker population and lower the entrepreneurship rate.

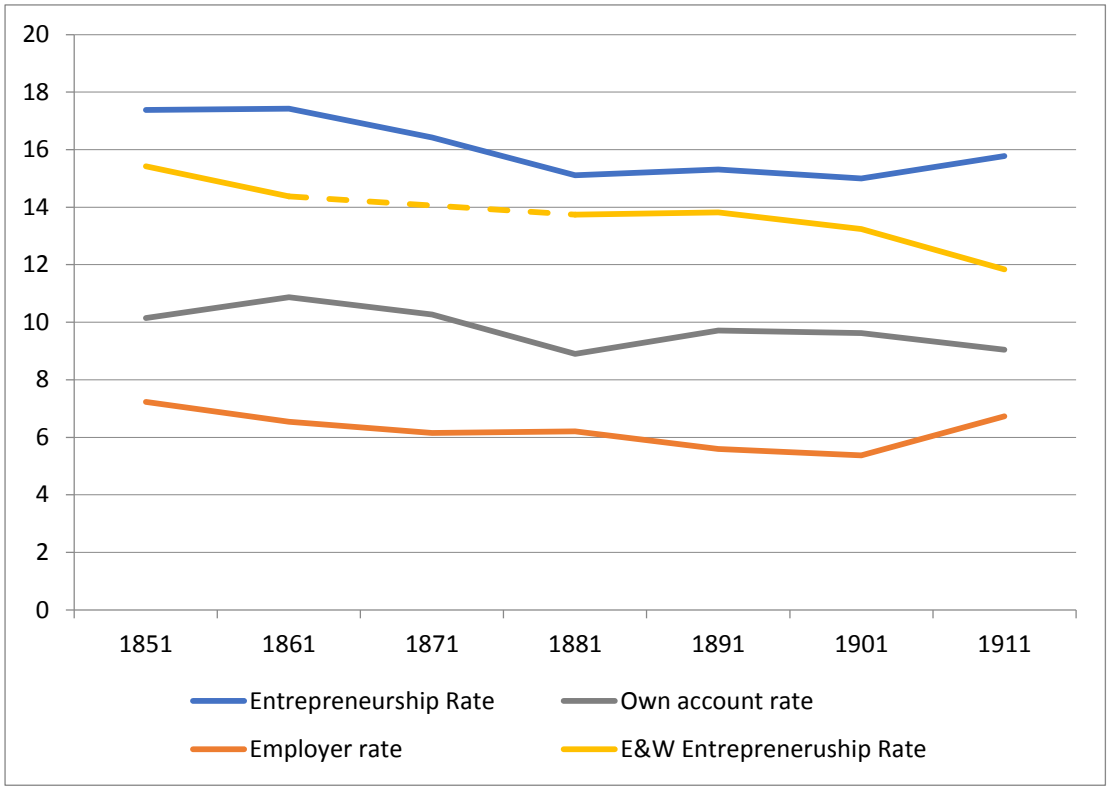


Figure 4. Trends in total entrepreneurship, employer, and own account entrepreneurship rates in Scotland, 1851-1911, weighted for missing 1861 farming data; no full England and Wales data for 1871.

In Scotland, as in England and Wales, the drop in entrepreneur numbers and rates between 1901 and 1911 was mainly driven by developments in the maker-dealer sector, as Figure 5 shows. Trends in most sectors can be examined using the EA17 aggregations as in England and Wales (see WP 5). The sector trends were similar to England and Wales: retail, transport, professional services, finance and commerce, agricultural produce and food sales all saw growth over this period, albeit at different rates, in common with England and Wales. Construction was more volatile than in England and Wales, reflecting the fact that the Scottish building industry was particularly volatile (Rodger, 1986). Mining peaked in 1901. A larger proportion of the Scottish mining industry was characterised by small-scale mining and it is possible that 1901 was a high point of this type of mining activity, before it began to be eroded in the twentieth century (Campbell, 2000). Refreshment was generally stable, but

this disguised changes within the sector, where whisky production was concentrating into fewer firms, but coffee houses, restaurants and other such small businesses were proliferating. Maker-dealers were relatively stable before a downturn after 1901, in contrast in England and Wales where maker-dealers rose steadily before falling after 1901. The difference likely arises from the fact that Scottish population growth was slower and thus the existing businesses were able to cater to the extra demand generated by population growth and thus there were fewer opportunities for new businesses to emerge. It is probable that the fall after 1901 is caused by similar factor in Scotland as in England and Wales, namely mechanisation, growing integration of the market for household goods, and increased concentration of the sector in larger firms. Manufacturing saw an acceleration of growth after 1891, suggesting that the turn of the century witnessed a significant expansion of the manufacturing sector both among employers and own-account proprietors.

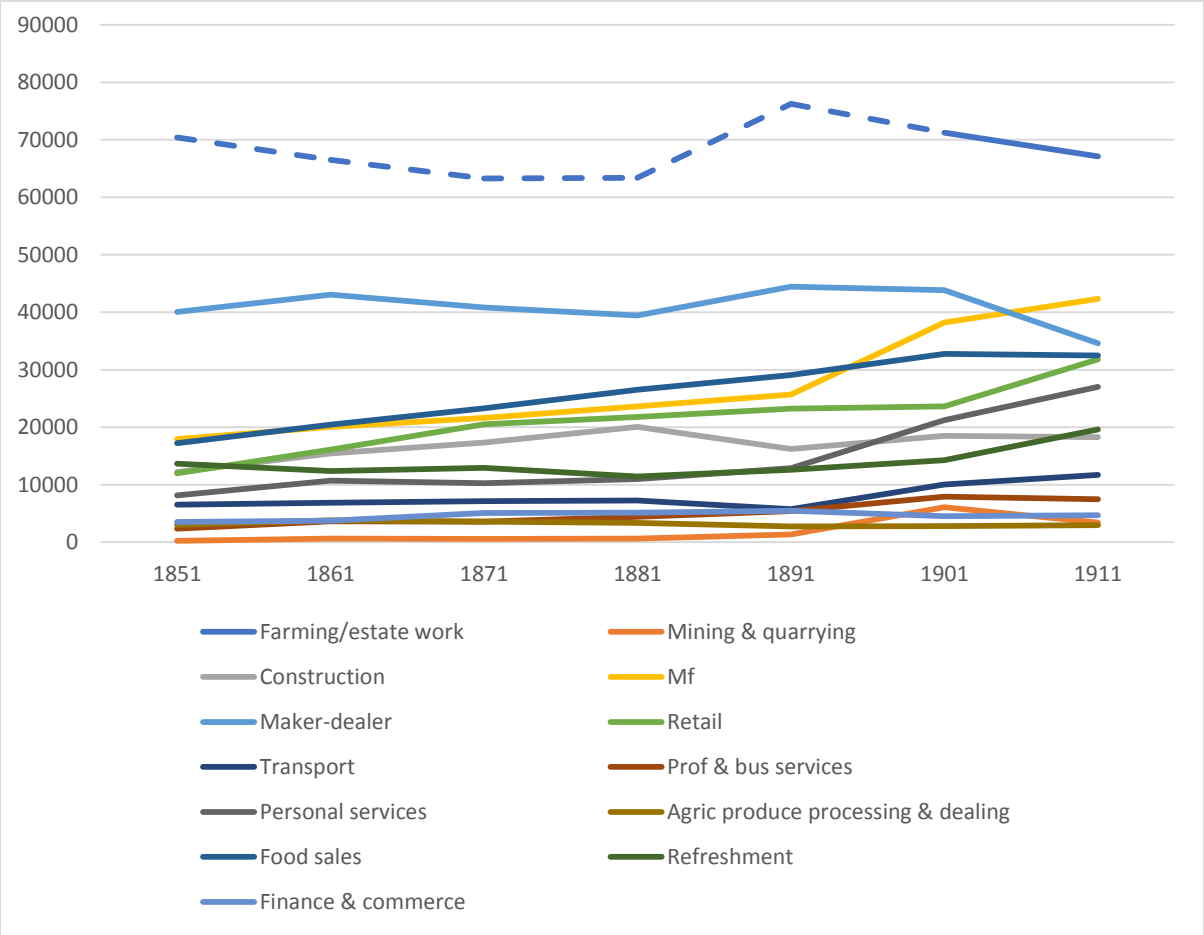


Figure 5. Number of entrepreneurs by EA17 sector, Scotland, 1851-1911; weighted for 1861 missing farming data; some uncertainty about number of farmers for 1851-91, as discussed in the text, shown dashed.

The trend in farming closely follows the aggregate published census reports, but now provides the full data for entrepreneurs for the whole period not previously available. Like England and Wales, there was a slowdown in the number of farm proprietors over the early period, reaching the lowest numbers in 1881, but the small decline 1851-61 (even after reweighting in Figure 5 for missing farm data) is the reverse of the small increase that occurred in England and Wales. Scotland's farm proprietors increased again to reach an all-time peak in 1891, but then declined steadily, which is a reversal of the continued increase up to 1911 experienced in England and Wales. The estimates for 1861 farmers are weighted and hence approximate (as noted in Section 2.4), and the height of the 1891 peak is uncertain. Hence, most of the farming data are shown dashed. The 1891 peak may result from the reconstruction model assigning too many crofters to worker status for 1881 and earlier. For example, crofters of worker status were in 1891 13 per cent, in 1901 17 per cent and in 1881 64 per cent, which suggests the earlier census supplementation for crofting is too low; and this is confirmed as a likely deficiency in our estimates by the more detailed work of Anderson and Roughley on these data. Some underestimation is also likely for 1871 and 1861 which all had a similar way of handling the farm/croft coding by the Scottish GRO (see Section 2.4). The data require further research. However, the total numbers involved in any estimation are small, and some of the 1891 peaking is likely to be real as a result of significant changes in Scottish farming and crofting in this period (see Campbell and Devine, 1990; Devine, 2018; Carter, 2001; Anderson, 2018). Improved interpretation of the farming and crofting data will be possible after completion of the detailed ongoing analysis by Anderson and Roughley. The calculations of the 1911 data for EA17 aggregates are given in Table 13 using GRO publications following the methods discussed in Section 4.

7. Business partners, directors, and portfolios in Scotland

Partners can be identified in the census records in many cases because an explicit statement such as 'partner of' is given. This accords with the census instructions that senior partners should make the return of the employee information of their firm, and conversely an implicit instruction that other partners should coordinate with the senior partner and not give the information as well. This was part of the instructions from 1861 until 1881, though relied on individual initiative to include such statements in 1851 and from 1891 onwards. In addition,

some respondents recorded that they were 'joint' proprietors with others[24] (especially in farming and crofting), sometimes giving the name or relationship to the other individuals concerned. Yet others gave implicit information about some sort of joint of partnership operations by stating they were 'operating with', 'working with', or just 'with'. These have been identified and coded to various forms of partnerships relationships using the same methodology applied in England and Wales. In addition, a larger proportion in Scotland was recorded as 'assisting with', which is an ambiguous statement that may relate to partnership. More details on the identification and coding of partners is given in WP 18, which extends the methods of extracting partnership census data for England and Wales in 1881 piloted in Bennett (2016). The coded partners of various forms are included in BBCE.

Some directors can be identified from the census and are available in BBCE from two sources: first, those who explicitly stated themselves as directors of companies in the census; and second those who are listed in the systematic coverage in the *Directory of Directors* who can be linked with their census entries (see WP 14). Linkage of Scottish directors achieved a similar rate as for England and Wales, with 36 per cent of the Directory 1881 directors linked to the census, 34 per cent in 1891, and 37 per cent in 1901. The methods for accomplishing this and coding the resulting individuals and their companies are described in WP 14. This was undertaken as one uniform process for England and Wales, and Scotland, with directors linked across the border where possible.

No attempt was made to take account of specific 'partner' or 'director' information in the reconstruction stage of the data construction in Scotland (the same process as in England and Wales). They were all treated as separate proprietors, estimated as employers or own account through the rest of the reconstruction. If the true status of directors in I-CeM coding was as a worker or not occupied they were identified subsequently by record-linkage with director directories (see WP 14). 'Partners' and 'directors' are, however, fully included in the BBCE data deposit.

[24] It should be noted that 'joint' could differ from its normal meaning in England and Wales by being applied to two different forms of activity on Scotland. It could mean two people sharing a croft/farm as de facto partners. But it could also mean sharing between several individuals or families some arable land (e.g. by 'rotation'); this was a very common feature in many parts of Scotland even in the early 19th century, but was target for reforming landlords, so it had become rare by the 1860s, though still practiced in a few areas. We are grateful to Mike Anderson for pointing out this distinction. The different types cannot be distinguished in the CEB entries. Each can be validly treated as a form of de facto or 'implicit' partnership and hence is broadly equivalent to the coding of 'joint' used in England and Wales. This is how they are coded in the BBCE, but further research at a local level might be able to add more detail to these individual BBCE codes.

Portfolios can be identified in the census from the occupational descriptors that responded to the census instruction to list activities in order of importance (for 1851-1901), and for 1911 giving 'main' occupation first. Portfolios were identified using a development of the method of Woollard (2004); see Bennett et al., (2019a, chapter 11). Occupational strings were searched for 'and', '&', 'also' and '+', the strings split into their component parts using the position of the word or symbol, and individual parts given occupation codes, with various types of false positives removed. This search was applied to the extracted entrepreneurs (employers and masters: Groups 1-6) in each year for 1851-81 with subsequent hand-coding, but in 1891-1901 the string search was applied to all the economically active population, with most splitting done automatically. Multiple non-entrepreneurial activities were excluded as covering by-employment rather than entrepreneurship, using the distinctions between categories defined in WP 3 and WP 5. Clerical corrections were then made to re-order portfolios so that the presumed primary business was listed first; this as far as possible matched the instruction in the census to record the main activity first, as noted above. This primarily aimed at correcting I-CeM coding that had picked up only part of a string which was a secondary activity, and also because I-CeM prioritised matching the criteria of census publications; for example, so that farmer was always identified and coded first, even if it was only a minor part of a portfolio with other businesses.

8. Urban classification using the Scottish Burghs

For many avenues of subsequent analysis it is valuable to be able to classify the types of locations where people were resident. For England and Wales an urban-rural classification was developed for BBCE based on the Law (1967) and Robson (1973) urban classification (see WP 6); but Law and Robson did not develop their classification for Scotland. Instead of trying to replicate what a Scottish Law-Robson classification might look like, a classification based on Scottish Burghs was used. The burghs have long been used by historians as useful units of analysis (Rodger, 1988; Morris, 1990), but their use here will make the data on urban areas a little different between England and Wales, and Scotland.

In all the 1851-1901 Census Reports there are lists of the Parliamentary Burghs and descriptions of the parishes which made up those Burghs.[25] In order to obtain an urban classification for Scotland, these lists were used to identify the I-CeM parishes which constitute these burghs.[26] For example, in 1851 Aberdeen covered all of the parishes of St Clement, North, Greyfriars, West, East and South, as well as part of the Parish of Banchory-Devenick and part of Old Machar. Each of these parishes can be identified in the I-CeM 1851 Scottish parish dictionary and coded to Aberdeen. The process was then repeated for all Burghs in each year. The fit is imperfect as it is impossible without further information to code only part of a parish to an urban unit. Consequently, to avoid attributing too many people to a town, if the population contributed by a parish to a Burgh was fewer than 250 then that parish was not coded to the Burgh. In the example of Aberdeen the parish of Old Machar included 30,503 people who were resident in Aberdeen and was coded to the Aberdeen urban unit, but in the parish of Banchory-Devenick only 5 people were living within the Parliamentary Burgh of Aberdeen, so that parish was not coded to Aberdeen. Where a parish included parts of more than one Burgh it was assigned to the larger Burgh; thus, the parishes of South Leith and St Cuthbert covered parts of both Edinburgh and Leith in each census between 1851 and 1891, they were all attributed to Edinburgh as that was the larger Burgh. The only exceptions to this were when the parish in question was the sole constituent parish of one of the Burghs in question. For example, the parish of Kinghorn covers part of both Kinghorn Burgh and Kirkcaldy Burgh in every census between 1851 and 1901. However, while Kirkcaldy Burgh has two other constituent parishes (Abbotshall and Kirkcaldy parishes), Kinghorn Burgh has none; thus Kinghorn parish was coded to Kinghorn Burgh.

This process creates a lookup table between I-CeM parish IDs and Parliamentary Burghs which can then be used to extract the populations of those Burghs for analysis (with the

[25] There are multiple definitions of the Scottish Burghs, based on different local government and political definitions. The census in this period consistently listed three: Parliamentary, Royal, and Police. Definitions of the constituent parts were only given for Parliamentary and Royal Burghs, so Police Burghs could not be used. Parliamentary Burghs were chosen over Royal because there were more of them in each census year.

[26] The lists are found in *Census of Great Britain, 1851, Population table 1, Vol. II, Parliamentary Papers*, LXXXVI (1852-3), 98-101; *Census of Scotland, 1861, Population tables and report*, *Parliamentary Papers*, L (1862), 152-3; *Census of Scotland, 1871, Eighth decennial census of the population of Scotland taken 3rd April 1871, with report, Vol. 1, Parliamentary Papers*, LXVIII (1872), 160-1; *Census of Scotland, 1881, Ninth decennial census of the population of Scotland taken 4th April 1881, with report, Vol. 1, Parliamentary Papers*, LXXVI (1882), 166-7; *Census of Scotland, 1891, Tenth decennial census of the population of Scotland taken 5th April 1891, with report, Vol. 1, Parliamentary Papers*, XCIV (1892), 172-3; *Census of Scotland, 1901, Parliamentary Burghs, districts of burghs and counties in Scotland*, *Parliamentary Papers*, CXXIX (1902), 2-3.

above caveats). Table 16 gives the urban economically active population, the number of entrepreneurs, the entrepreneurship rate and the share of all entrepreneurs present in Parliamentary Burghs for each census year using only the Burghs present in each census year as the list of Burghs changes somewhat year to year. More can be done, but this tabulation shows the importance of towns to Scottish entrepreneurship, an importance which is even starker if farming is removed from the calculations.

	1851	**1861**	**1871**	**1881**	**1891**	**1901**
Economically Active	495,644	546,984	719,582	849,582	953,181	1,132,475
Entrepreneurs	76,079	87,234	98,309	103,050	115,141	138,181
Entrepreneurship Rate	15.3	15.9	13.7	12.1	12.1	12.2
Share of Entrepreneurs	36.7	39.8	42.7	43.3	44.1	46.8
Share of Employers	30.3	36.4	41.8	41.6	42.5	43.5
Share of Own Account	41.2	41.8	43.3	44.4	45.0	48.7

Table 16. Entrepreneur characteristics of Parliamentary Burghs, 1851-1901.

9. Conclusion

This Working Paper has outlined how the census of the population in Scotland can be used to identify businesses proprietors, and where the census needs to be supplemented by other sources and estimation. The paper has outlined how the census allows employers, own-account proprietors, and workers to be identified and/or estimated. It also outlines how partnerships, directors and portfolio businesses can be identified. The result is a database of entrepreneurs for Scotland contained in BBCE, which can be linked with I-CeM to give estimates that are as complete as possible of the census coverage of Scottish business people and their demographic characteristics. This leads to the important conclusion that Scottish entrepreneurship rates were higher than England and Wales; and these are probably slightly underestimated as a result of the lower levels of own account and employer activity in crofting etc. that are used in this paper. For both Scotland and England and Wales there is also a slight underestimation in BBCE because of the exclusion of the records of ‘xxx wife’.

The full coding resulting from the Scottish efforts discussed here is included in the BBCE. Additionally, downloads of the reconstruction decision process with this Woking Paper 20; Cut-off information for farms by parish, and the estimation of the Workers in the whole active population with their occodes are also given with the WP 20; the Scottish Burghs classification is given in a download with WP 6; downloads of alternative reconstruction model estimates and a guide to use estimation process are provided and with WP 19.

It is acknowledged that the population census was not designed as a business census; hence the information gathered by census administrators constrains the business information that can now be obtained and its level of completeness. For the supplemented data the reconstructed estimates are provided as statistical estimates that infill otherwise missing census reposes. But users should recognise that the supplemented data provided for individuals are *estimates* that are no more than a best statistical calculation. *The supplementation method cannot produce perfect estimates. Users of the BBCE data deposit listed as EMPSTATUS_IND must treat this with extreme care, especially for small parishes and small sub-categories of the population, and regard the data as preliminary estimates for farmers and crofters in the Highlands and Islands.*

However, despite caveats, the Scottish census fills many of the statistical gaps for business proprietorship that have previously impeded research. As a source of systematic information on most employers and own account self-employed, and for early years also the size of the workforces of employers and the acreage of farms, it is unparalleled and allows the large-scale study of entrepreneurship for all of Great Britain between 1851 and 1911. This paper describes how the information for Scotland has been brought up to the level available in BBCE for England and Wales on an aligned and fully compatible basis.

Downloads of the background decisions and calculations behind the analysis and BBCE database for Scotland are linked to this paper; equivalent downloads for England and Wales accompany the equivalent WPs, as listed at the end of this paper, The downloads for Scotland are:

- Decisions on reconstruction estimates for each of the tailored logit cut-offs 1851-81
- The cut-offs by parish for the farm estimates 1851-81

- The full economically active by RecID, for corrected Occodes and Sub-Occode for Scotland

Acknowledgments:

This research has been supported by the ESRC using a special supplementation of the ESRC-supported project ES/M010953 **Drivers of Entrepreneurship and Small Businesses** to include Scotland so that full aggregates of Britain were available in the database deposit of the *British Business Census of Entrepreneurs* (BBCE). An Isaac Newton Trust Grant, 18.40(g) **'Business proprietor succession and firm size change 1851-1881'**, provided additional valuable support to help track and code Scottish employer and director data. Tobias Lund, Annette Mackenzie, and Amber Stevenson were supported as research assistants in Scotland from these funds; we are very grateful for their work.

The Scottish census database used derives from Higgs, Edward and Schürer, Kevin (University of Essex) (2014) *The Integrated Census Microdata (I-CeM)* UKDA, SN-7481; see also E. Higgs, C. Jones, K. Schürer and A. Wilkinson, *Integrated Census Microdata (I-CeM) Guide*, 2nd ed. (Colchester: Department of History, University of Essex, 2015).

A special acknowledgement of thanks is made to Kevin Schürer for advice and help in providing advanced issue of some additional Scottish material that will be in updated versions of I-CeM.

We are also extremely grateful to Michael Anderson and Corinne Roughley for improving the Scottish parish PARID and COMPARID files and dictionaries, for identifying duplicate and spatially misallocated records for farmers, crofters etc., for comments on earlier versions of this WP, and for other advice, particularly on farmers and crofters. However, the data in BBCE do not include all the further improvements they have made to identifying crofters, tenants and other categories which form part of their subsequent analysis.

The GIS boundary files for parishes were derived from Corinne Roughley (2019a) *Scotland's Parish Populations: parish boundaries 1755-1891*, and Corinne Roughley (2019b) *Scotland's Parish Populations: parish boundaries 1891-1911*, deposited at National Records of Scotland.

References

Anderson, Michael (2018) *Scotland's Populations from the 1850s to Today*, Oxford: Oxford University Press.

Bennett, Robert J. (2016) 'Interpreting business partnerships in late-Victorian Britain', *Economic History Review*, 69 (4), 1199-1227.

Bennett, Robert J. and Newton Gill (2015) Employers and the 1881 population census of England and Wales, *Local Population Studies*, 94, 29-49.

Bennett, Robert J., Smith, Harry, van Lieshout, Carry, Montebruno, Piero and Newton, Gill (2019a) *The Age of Entrepreneurship: Business Proprietors, Self-employment and Corporations Since 1851*, Abingdon: Routledge.

Bennett, R. J., Smith, H., Montebruno, P. (2019b) 'The Population of Non-corporate Business Proprietors in England and Wales 1891–1911', *Business History*,doi.org/10.1080/00076791.2018.1534959

Bennett, Robert J., Smith, Harry, Montebruno, Piero, and van Lieshout, Carry (2020) 'Changes in Victorian entrepreneurship in England and Wales, proprietor numbers, rates, and gender, 1851-1911', forthcoming

Campbell, Alan (2000) *The Scottish Miners, 1874-1939: Industry, Work and Community*, 2 vols., Aldershot: Ashgate.

Campbell, R.H. and Devine, T.M. (1990), 'The Rural Experience', in W. Hamish Fraser and R.J. Morris (eds), *People and Society in Scotland: Volume II, 1830-1914*, Edinburgh: John Donald, 46-72.

Carter, Ian R. (2001), *Farmlife in Northeast Scotland, 1840-1914*, Edinburgh: John Donald.

Devine, T.M. (2018), *The Scottish Clearances: A History of the Dispossessed*, London: Allen Lane. Higgs, Edward (2005) *Making Sense of the Census Revisited: Census Records for England and Wales 1801-1901*, London: Institute of Historical Research and National Archives.

Higgs, Edward and Schürer, Kevin (2014) *The Integrated Census Microdata (I-CeM)* UK Data Service, SN-7481 [data collection]. http://dx.doi.org/10.5255/UKDA-SN-7481-1

Higgs, E., Jones, C., Schürer, K. and Wilkinson, A. (2015), *Integrated Census Microdata (I-CeM) Guide*, Colchester: Department of History, University of Essex.

Law, Chris M. (1967) 'The Growth of Urban Population in England and Wales, 1801-1911', *Transactions of the Institute of British Geographers*, 41, 125-43.

Montebruno, P., Bennett, R.J., van Lieshout, C., Smith, H. and Satchell, M. (2019) Shifts in agrarian entrepreneurship in mid-Victorian England and Wales, *Agricultural History*

Review, 67, 1, 71-104.https://www.ingentaconnect.com/content/bahs/agrev/2019/00000067/00000001/art00006

Morris, R.J. (1990), 'Urbanisation and Scotland', in W. Hamish Fraser and R.J. Morris (eds), *People and Society in Scotland: Volume II, 1830-1914*, Edinburgh: John Donald, 73-102.

Moss, Michael S. and Hume, John R. (1983), 'Business Failure in Scotland, 1837-1913: A Research note', *Business History*, 25/1, 3-10.

Raffo, Julio (2015) *MATCHIT: Stata module to match two datasets based on similar text patterns*, Statistical Software Components S457992, Boston College Department of Economics, revised 09 Apr 2017.

Robson, Brian (1973) *Urban Growth: An Approach*, London: Methuen.

Rodger, Richard, 'Business Failure in Scotland, 1837-1913', *Business History*, 27/1 (1985), 75-99.

Rodger, Richard (1986) 'Structural instability in the Scottish building industry, 1820-80', *Construction History*, 2, 48-60.

Rodger, Richard (1988) 'Concentration and Fragmentation: Capital, Labour, and the Structure of Mid-Victorian Scottish Industry', *Journal of Urban History*, 14, 2, 178-213.

Roughley, Corinne (2019a) *Scotland's Parish Populations: parish boundaries 1755-1891*, deposited at National Records of Scotland.

Roughley, Corinne (2019b) *Scotland's Parish Populations: parish boundaries 1891-1911*, deposited at National Records of Scotland.

Schürer, K. (1991) 'The 1891 Census and Local Population Studies', *Local Population Studies*, 47, 16-29.

Schürer, K., Higgs, E., Reid, A.M. and Garrett, E.M. (2016) *Integrated Census Microdata, 1851-1911, User Guide version V. 2 (I-CeM.2)*, Second edition, Colchester: Department of History, University of Essex.

Woollard, Matthew (2004) 'The Classification of Multiple Occupational Titles in the 1881 Census of England and Wales', *Local Population Stu*dies, 72, 34-49.

Other Working Papers:

Working paper series: ESRC project ES/M010953: *'Drivers of Entrepreneurship and Small Business',* University of Cambridge, Department of Geography and Cambridge Group for the History of Population and Social Structure.

WP 1: Bennett, Robert J., Smith Harry J., van Lieshout, Carry, and Newton, Gill (2017) *Drivers of Entrepreneurship and Small Businesses: Project overview and database design.* https://doi.org/10.17863/CAM.9508

WP 2: Bennett, Robert J., Smith Harry J. and van Lieshout, Carry (2017) *Employers and the self-employed in the censuses 1851-1911: The census as a source for identifying entrepreneurs, business numbers and size distribution.* https://doi.org/10.17863/CAM.9640

WP 3: van Lieshout, Carry, Bennett, Robert J., Smith, Harry J. and Newton, Gill (2017) *Identifying businesses and entrepreneurs in the Censuses 1851-1881*. https://doi.org/10.17863/CAM.9639

WP 4: Smith, Harry J., Bennett, Robert J., and van Lieshout, Carry (2017) *Extracting entrepreneurs from the Censuses, 1891-1911*. https://doi.org/10.17863/CAM.9638

WP 5: Bennett, Robert J., Smith Harry J., van Lieshout, Carry, and Newton, Gill (2017) *Business sectors, occupations and aggregations of census data 1851-1911*. https://doi.org/10.17863/CAM.9874

Data download of classification file: https://doi.org/10.17863/CAM.9874

WP 6: Smith, Harry J. and Bennett, Robert J. (2017) *Urban-Rural Classification using Census data, 1851-1911*. https://doi.org/10.17863/CAM.15763

WP 7: Smith, Harry, Bennett, Robert J., and Radicic, Dragana (2017) *Classification of towns in 1891 using factor analysis.* https://doi.org/10.17863/CAM.15767

WP 8: Bennett, Robert J., Smith, Harry, and Radicic, Dragana (2017) *Classification of occupations for economically active: Factor analysis of Registration Sub-Districts (RSDs) in 1891*. https://doi.org/10.17863/CAM.15764

WP 9: Bennett, Robert, J., Montebruno, Piero, Smith, Harry, and van Lieshout, Carry (2018) *Reconstructing entrepreneurship and business numbers for censuses 1851-81*. https://doi.org/10.17863/CAM.37738

WP 9.2: Bennett, Robert, J., Montebruno, Piero, Smith, Harry, and van Lieshout, Carry (2019) *Reconstructing business proprietor responses for censuses 1851-81: a tailored logit cut-off method.*

WP 10: Bennett, Robert, J., Smith, Harry and Radicic, Dragana (2018) *Classification of environments of entrepreneurship: Factor analysis of Registration Sub-Districts (RSDs) in 1891*. https://doi.org/10.17863/CAM.26386

WP 11: Montebruno, Piero (2018) *Adjustment Weights 1891-1911: Weights to adjust entrepreneur numbers for non-response and misallocation bias in Censuses 1891-1911*. https://doi.org/10.17863/CAM.26378
Adjustment weights: https://doi.org/10.17863/CAM.26376

WP 12: van Lieshout, Carry, Day, Joseph, Montebruno, Piero and Bennett Robert J. (2018) *Extraction of data on Entrepreneurs from the 1871 Census to supplement I-CeM.* https://doi.org/10.17863/CAM.27488

WP 13: van Lieshout, Carry, Bennett, Robert J. and Smith Harry (2019) *Extracted data on employers and farmers compared with published tables in the Census General Reports, 1851-1881*. https://doi.org/10.17863/CAM.37165

WP 14: van Lieshout, Carry, Bennett Robert J. and Montebruno, Piero (2019) *Company Directors: Directory and Census record linkage.* https://doi.org/10.17863/CAM.37166

WP 15: Bennett, Robert, J., Montebruno, Piero, Smith, Harry and van Lieshout, Carry (2019) *Entrepreneurial discrete choice: Modelling decisions between self-employment, employer and worker status.* https://doi.org/10.17863/CAM.37312

WP 16: Satchell, M., Bennett, Robert J., Bogart, D. and Shaw-Taylor, L. (2019) *Constructing Parish-level Data and RSD-level Data on Transport Infrastructure in England and Wales 1851-1911*. https://doi.org/10.17863/CAM.37313

WP 17: Satchell, M. and Bennett, Robert J. (2019) *Building a 1911 Historical Land Capacity GIS*. https://doi.org/10.17863/CAM.42285

WP 18: Bennett, Robert, J., Smith, Harry, van Lieshout, Carry and Montebruno, Piero (2019) *Identification of business partnerships in the British population censuses 1851-1911 for BBCE.* https://doi.org/10.17863/CAM.43890

WP 18: Bennett, Robert, J., Smith, Harry, van Lieshout, Carry and Montebruno, Piero (2019) *Identification of business partnerships in the British population censuses 1851-1911 for BBCE.* https://doi.org/10.17863/CAM.43890

WP 19: Montebruno, Piero (2019) *Datasets and guide: downloads for reconstructing British census responses 1851-1881 for the BBCE*. https://doi.org/10.17863/CAM.42285
Includes linked downloads of files by RecID for all Scotland, and England and Wales for all economically active 1851-81: employers, own account, and workers.

WP 20: Smith, Harry, van Lieshout, Carry, Montebruno, Piero and Bennett, Robert, J. (2019) *Preparing Scottish census data in I-CeM for the British Business Census of Entrepreneurs (BBCE)*.

Full list of all current Working Papers available at:
http://www.geog.cam.ac.uk/research/projects/driversofentrepreneurship